Comparing Supervised Classification Methods on a Multispecies Fishery of Juvenile Fish:

Galaxiid Whitebait Classification

A thesis submitted in partial fulfilment

of the requirements for the degree of

Masters of Science in Statistics

By Bridget Ann Armstrong

Under the supervision of:

Dr Elena Moltchanova,

Dr Mike Hickford

And

Dr David Schiel

School of Mathematics and Statistics

University of Canterbury

2019

Contents:

List of Figures

observations are indicated in red.

List of Tables

Table 33 Selected confusion matrices for Random Forest. All percentages are CV error. In national data, giant kōkopu were misclassified 24.79% of the time as kōaro. In Marlborough, kōaro were frequently misclassified as īnanga. In Otago 93.55% of kōaro were misclassified as īnanga. Giant kōkopu in Tasman and Wellington were all misclassified. In Tasman giant kōkopu were misclassified as kōaro, and in Wellington giant kōkopu were either misclassified as banded kōkopu or kōaro. pg 94

Acknowledgements

The data set I used was compiled by Mark Yungnickel as part of his Master's thesis. I thank Mark greatly for what must have been a long and difficult task. You have made cleaning the data much easier than it could have been, and provided me with insight at the earlier stage of data analysis. It is with great respect that I thank the whitebaiters who have contributed their samples.

To my supervisors; I am in awe at your compassion, patience, and ability to keep me going through the most difficult of times. Thank you all for everything.

Without everybody's contribution I would not be able to have studied in my favourite two fields; statistics and ecology.

“One man’s fish is another man’s poisson” – Anon

Glossary

Amphidromy: a type of diadromy where fish are born in fresh or brackish water, then drift into the ocean as larvae before migrating back into freshwater to grow into adults and spawn.

Banded kōkopu*: Galaxias fasciatus*

Brackish water: Water with between 0.5 to 30 grams of salt per litre (Remane and Schlieper 1972)

Bycatch: other non-target species of fish that are caught with the target species

CV: Cross validation

Diadromous/diadromy: A lifestyle of aquatic animals that means spawning freshwater, larvae hatch and spend part of their life in the ocean, then return to freshwater to mature

Estuary: The tidal opening of a freshwater body to the ocean

Giant kōkopu*: Galaxias argentus*

Īnanga*: Galaxias maculatus,* or īnaka (Ngāi Tahu/South Island iwi)

Kōaro*: Galaxias brevipennis*

LOO: Leave out one cross validation

Metapopulation: a population of animals that have weaker genetic connections between spatially divided groups

Natal stream: the stream an animal was birthed or hatched from

Otolith: ear bones

Shortjaw kōkopu: *Galaxias postvectis*

Whitebaiter: someone who fishes for whitebait

Whitebaiting: fishing for whitebait

Section 1: Introduction

Post-larvae (whitebait) of the genus Galaxias constitute an iconic fishery in New Zealand. The fishery is essentially nationwide but is largely concentrated on the West Coast of the South Island. It is known that there are wide geographic differences in morphometrics of the whitebait catch, but this is complicated by the fact that five Galaxias species constitute the catch. The five species are īnanga (*Galaxias maculatus*), kōaro (*Galaxias brevipinnis*), banded kōkopu (*Galaxias fasciatus*), giant kōkopu (*Galaxias argentus*), and shortjaw kōkopu (*Galaxias postvectis*). One of these five species, banded kōkopu, is relatively easy to identify because of its size and markings, and one species, shortjaw kōkopu, is especially rare and considered endangered. Around 88% of the whitebait catch is one species, called īnanga. Whitebait are caught in enormous abundances as they return from development in the oceanic environment to freshwater streams and rivers. Because these fish are small, post-larval juveniles it is difficult to tell the species apart to characterise the exploitation level of each species (Figure 1). Microscopic examination and, in the case of shortjaw kōkopu, genetic analyses can give identifications with high accuracy, but these techniques are expensive and time consuming. This work aims to use statistical classification to reliably distinguish species with morphological measures. It is a novel classification method for whitebait. These methods use data from extensive sampling of whitebait from 15 regions around New Zealand collected in another study (Yungnickel 2017). It is a data-rich source of morphological measurements by geographic region and is based on around 17500 observation and measurements. This thesis is laid out as follows: a background of the whitebait fishery, life histories and current knowledge; a description of each method with example data to demonstrate the differences in methods; a description of the data; results from the methods; followed by discussion and recommendations for future work.

Figure 1: Juvenile whitebait freshly caught from a river. There could be up to five different species of Galaxiid in this bucket. Adapted from Yungnickel 2017.

The Fishery

Whitebait are an important cultural icon in New Zealand (Baker Egan & Gee, 2018). Over the past few decades limited evidence has suggested that catches of whitebait have declined (Goodman, 2018). Knowing which species are declining requires consistent monitoring, and the ability to distinguish them. The adult fishes are remarkably unique physically and ecologically (McDowall 1964) but difficult to sample. The juveniles, or whitebait, are caught in high abundance but species are difficult to distinguish (Yungnickel, 2017). Biodiversity in ecosystems is like an insurance policy against system stressors (Elmqvist, Folke, Nyström, Peterson, Bengtssorn, Walker & Norberg, 2003). That is, if a system is more diverse in terms of species, it will be more resilient in response to stressors.

In this thesis I have studied the classification of whitebait, a multispecies fishery. It is known that there are wide geographic differences in the morphometrics of the whitebait catch, but this is complicated by the fact that five Galaxias species constitute the catch. Whitebait is defined under Section (2) of the *Whitebait Fishing Regulations 1994* and Section (2) of the *Whitebait Fishing (West Coast) Regulations 1994* as 'the young or fry of īnanga (*Galaxias maculatus*), kōaro (*G. brevipinnis*), banded kōkopu (*G. fasciatus*), giant kōkopu (*G. argenteus*), shortjaw kōkopu (*G. postvectis*), and common smelt (*Retropinna retropinna*)'. However, the migratory galaxiids (īnanga, kōaro, banded kōkopu, giant kōkopu and shortjaw kōkopu) are considered 'true' whitebait by both biologists and

whitebaiters. The sixth whitebait species, *Retropinna retropinna* (smelt) is included as whitebait under fishing regulations but is not investigated here as they are easily distinguished by their unique smell (Yungnickel, 2017). The fishing season occurs between August 15th and 30th November of each year in the North and South Island of New Zealand, and 1st September to 30th November on the West Coast of the South Island using set nets and scoop nets (*Whitebait Fishing Rgeulations 1994)*. During this time an estimated 200 tonnes of whitebait are landed (Yungnickel, 2017). Estimations of annual whitebait capture is from sparse records as the data is very difficult to get (Goodman, 2018). Whitebait are small fish (Figure 1), usually less than 0.5g, and less than 60mm long (Table 1) (McDowall, 1964)so an individual whitebaiter's catch can prevent about 10,000 individuals that would potentially recruiting into populations. At present, the fishery is seldom surveyed as it is expensive (Goodman, 2018). As such, current fishery management relies on regulations that were developed over 20 years ago(Baker, Egan, & Gee, 2018; McDowall 1996). The long term sustainability of the fishery is not well understood (Goodman, 2018).

Fish Life History

Galaxiid whitebait species have a similar life history although, details are uncertain for some species (Goodman, 2018). In general, adults spawn in fresh to brackish water newly hatched larvae are washed out to sea by tidal or flood inundation (McDowall 1970). The larvae spend approximately three in the marine environment before they return to freshwater as post-larvae; this lifestyle is known as diadromy, specifically amphidromy. Some landlocked populations also occur (Goodman, 2018), although they have not been investigated here. What is known is that the life history of each species is subtly different and the proportion of each species in the overall whitebait catch is not the same in each region, or across the season (Egan, 2017; Rowe & Kelly, 2009; Yungnickel, 2017). Species are superficially similar at the whitebait stage (Figure 1) but grow into physically (Figure 2) and ecologically distinctive fish (Table 1). Īnanga are the shortest lived and smallest and tend to reproduce once. All other species reproduce more than once. Giant kōkopu is longest at maturity but does not have the longest whitebait (Table 1). Length ranges for whitebait by species have been collated and summarised in Table 1 (Woods, 1968; Egan, 2017; and Yungnickel, 2017).

Table 1: The life history traits and selected metrics of each species. Species are superficially similar at the whitebait stage with similar lengths but grow into distinctive fish. Īnanga are the shortest lived and shortest. Īnanga tend to reproduce once. All other species reproduce more than once. Giant kōkopu is longest at maturity but does not have the longest whitebait. Shortjaw kōkopu are the rarest as adults and are rarely seen as whitebait.

Species	Life expectancy (yrs)	Lifetime spawning events	Percentage of whitebait catch	Length range whitebait (mm)	Length range mature (mm)
Īnanga	1	1	88.20	36.4 – 59.6	80-110
Kōaro	~ 10	6-8	5.00	36.2 – 59.7	160-180
Banded Kōkopu	~ 10	>1	6.60	33.8 – 48.5	200
Giant Kōkopu	>25	>1	0.03	34.0 – 55.4	300-450
Shortjaw Kōkopu	~ 10	>1	0.01	40.3 – 57.9	150-200

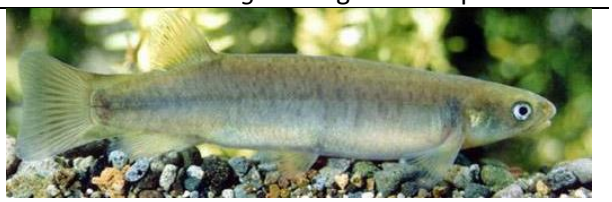
Studies About Whitebait

Whitebait research often focuses on ecology of selected species (Allibone & Caskey, 2000; Hickford & Schiel, 2010), conservation status and efforts (Allibone et al., 2010; Goodman et al., 2014), and fishery management (Baker, Egan, & Gee, 2018; McDowall, 1965). Other projects ask how species are distributed in time and space (Egan, 2017; Yungnickel, 2017), the abundance of each species in whitebait catches (Yungnickel, 2017), and how different species of galaxiid interact with predatory and introduced fish (Glova, 2003; McLean, Barbee, & Swearer, 2007). Knowledge about reproduction and lifespan are well documented for īnanga, but the remaining species are not well understood. The least common species, short jaw kōkopu is found in very low abundances and has a patchy distribution as whitebait and adults (Goodman et al., 2014). Recently, four out of the five Galaxiid whitebait species were classified as “at risk” or “in decline” (Dunn et al., 2018). Why adults of Galaxiid species may be in decline is a matter of wide speculation, although degradation of spawning sites and habitat destruction has been implicated (Orchard, Hickford, & Schiel, 2018; Taylor, 2002). To have a better understanding of how all pressures affect each species, more detailed knowledge is required about each species. At present the most is known about īnanga in terms of species ecology, species interactions, and fishery exploitation. Considerably less is known about the remaining species; particularly, knowledge about the fishery exploitation of these species is sparse (Goodman, 2018).

Galaxias maculatus: īnaka/īnanga

Galaxias argentus: giant kōkopu

Galaxias brevipennis: kōaro



Galaxias postvectis: shortjaw kōkopu

Galaxias fasciatus: banded kōkopu

Retropinna retropinna: common smelt

Figure 2: Adults of all whitebait species that are legally allowed to be fished as a part of the whitebait fishery according to Department of Conservation fishery regulations. Images by Stephen Moore, used with permission.

What is known is that each species is the whitebait migrate to freshwater from May to November (Woods, 1968), but each species peak migration timing varies (Goodman, 2018). The peak migration of īnanga tends to be from August to November; kōaro and banded kōkopu from September to October, giant kōkopu having peak migration estimated to be in November. Shortjaw kōkopu are rarely seen (Goodman, 2018). Given this variation in migration timing, there is also variation in the size of whitebait between the species, across species, and across a season (Goodman, 2018; Woods, 1968).

Whitebait Identification

Identification of the species of whitebait has been covered by few authors (Charteris & Ritchie, 2002; Dijkstra & McDowall, 1997; Woods, 1968; Yungnickel, 2017). As a result, identification methods of whitebait have remained largely unchanged for many years (Yungnickel, 2017; Woods, 1968). Projects to sample and identify whitebait at all stages of life that cover a large spatial scale are rare as they are expensive (Goodman, 2018). What has been discovered the species are different, but it requires patience and a microscope to tell the difference (Yungnickel, 2017).

Genetic identification is an alternative identification method, but requires considerable training and cost (Dijkstra & McDowall, 1997). Specimens used for identification are destroyed and are unable to

be checked unless there is more of the specimen to sample from. Dijkstra and McDowall identified specimens from a small region in the South Island of New Zealand using DNA (1997). However, this may not be adequate as there is evidence that the fishery of the country is poorly mixed (Goodman, 2018), the genetic identification of species for one area may not be adequate for all regions of New Zealand. Yungnickel (2017) used DNA classification to confirm the species of shortjaw kōkopu, although they were all from the same region.

What other methods are available?

Phenological matching and genetic techniques are both time-consuming and expensive. An alternative way of classifying species is by using statistical classification techniques using commonly obtained metrics of an animal. There are different types of statistical classification with different advantages. Here I use supervised classification. Supervised statistical classification for identifying animals often uses images or measurements from specimens. Models are produced by using labelled data to recognize differences between species. The model is then used to classify unseen observations into species. Various fish morphometric measures have been used to classify fish using supervised classification (D'Elia et al., 2014; Guisande et al., 2010; Jones & Checkley Jr, 2017). However, the models need to be built using reliably labelled data (D'Elia et al., 2014; Gaston & O'Neill, 2004; Guisande et al., 2010).

Supervised classification methods can be useful for reducing the cost of species identification for fish (Gaston & O'Neill, 2004; Guisande et al., 2010). Instead of transporting whitebait specimens to a lab having to identify them, they could be identified using simple body measures at the riverside. Until now, there have been no efforts to classify whitebait using statistical classification models. With statistical models, most require a small amount of training with the computer program that predicts the classifications. Once this has been achieved the data can be parsed to the model and classifications made for a specimen in seconds without destroying the sample, and without having to compare it other observations. Therefore the aims of my thesis are:

1. Use morphometric data to classify whitebait into species

2. Examine and categorise geographic differences within and among species

3. Compare the performance of five different classification methods for quick identification to examine catch composition by species.

It is envisioned that these sorts of results, if successful, could be developed into quick assays.

The structure of this Thesis is as follows. In Section 2 I will briefly explain the different statistical classification methods used in this study. I will then describe the data and the data collection methods in Section 3. In Section 4 I outline exploratory data analysis and describe the results. And finally, in Section 5, I will discuss the meaning of these results and potential for any future work.

Section 2: Methods

The purpose of this project was to cheaply and quickly classify the five species of whitebait morphometric measures with location and date measures. Supervised classification methods were chosen as the data was labelled with species. The five classification methods were multinomial logistic regression (MLR), linear discriminant analysis (LDA), Naïve Bayes, Decision Trees, and Random Forest. Quadratic discriminant analysis (QDA) was tried , however, some variance-covariance matrices for individual species within regions were singular, so QDA was abandoned (Hastie & Tibshirani, 2009; James, Witten, Hastie, & Tibshirani, 2013). Data pooled by regions (national data) were modelled to capture the national patterns of species morphological distributions, and data from each region (regional data) were modelled separately to capture spatial differences in species morphological distributions using methods described in this chapter.

Before discussing methods, I will explain how I assessed each model. The fit of each model was assessed by looking at the confusion matrices and comparing apparent error rates (AER). A confusion matrix is a kind of contingency table which counts the number of observations for each category (Table 2). Each observation is either correctly classified, or incorrectly classified. The count of correctly classified observations occurs on the diagonal where the observed and predicted species are the same. The incorrectly classified observations are on off-diagonals with the row giving the species that the observations were misclassified as. AER is the percentage of incorrectly classified individuals. The best models had the lowest AER (Agresti, 2013).

Table 2: Example confusion matrix. In total there are 25 specimens that are Class one, and 21 that are Class two. For Class one, 10 observations have been correctly classified, and 15 have been misclassified as Class two. For observations from Class two 20 have been correctly classified and one has been misclassified as Class one

	Observed Count Class One	Observed Count Class Two
Predicted Count Class One	10	1
Predicted Count Class Two	15	20

Cross validation

Cross validation (CV) has been designed to quantify the predictive performance of a model (Kohavi, 1995). There are two types of cross validation, k-fold and leave out one (LOOCV) (James et al., 2013). The key steps in cross validation are (James et al., 2013):

1. Leave out either one observation (LOOCV) or a set of observations (k-fold) which form a set of unseen data.
2. Recalculate the prediction model with the remaining set of observations, or labelled observations.
3. Predict the response variable and evaluate the associated prediction error for the unseen data based on the new model.
4. Repeat steps 1 – 3 until there is a prediction error for each set of unseen data.
5. Average the prediction errors over all the unseen data.

The LOO method is an exhaustive cross validation measure which means that all possible combinations of each observation being in the unseen set is explored. Due to the exhaustive search LOO can be computationally expensive (Kohavi, 1995).

K-fold cross validation is computationally less intensive than leave out one cross validation. It is a non-exhaustive cross validation method because not all combinations of k-folds are explored. Folds are assigned randomly to reduce bias. Ten folds are customarily used (Kohavi, 1995). Model accuracy is assessed by calculating the proportion of correctly identified observations for all species, and for each species (James et al., 2013).

Cross validation is used to assess over-fitting of a model. When a model is overfit the AER will be close to zero, but the CV error will be large (James et al., 2013).

Toy data

Table 3: Descriptive statistics of each toy data set. In toy data set one the species are mixed and have a mean for Northing and Length. There are more īnanga than giant kōkopu in this set. For toy data two, the species have slightly different means for Nothing and Length. Each species is present in equal quantities. For toy data three species have different means for Northing and similar means for Length. Toy data four species have similar means for both Length and Northing, but the standard deviation for giant kōkopu is large.

Set	Species	n	mean length (sd)	mean northing (sd)
1	īnanga	54	47.9 (1.03)	-41.2 (0.88)
	giant kōkopu	25	48.2 (0.92)	-40.9 (1.04)
2	īnanga	40	48.6 (0.81)	-40.3 (0.73)
	giant kōkopu	40	47.4 (0.78)	-41.7 (0.70)
3	inānga	38	48.6 (0.63)	-41.9 (0.46)
	giant kōkopu	41	47.4 (0.96)	-40.2 (0.61)
4	inānga	54	47.6 (0.77)	-41.2 (0.60)
	giant kōkopu	79	48.3 (1.06)	-40.9 (1.19)

To demonstrate each method, I simulated four toy datasets based on the 'length' and 'northing' variables for the species īnanga and giant kōkopu (Figure 3). Each data set has different properties to highlight the differences between the methods. Descriptive statistics of each toy data set are given in Table 3 with variance-covariance matrices in Table 4 with plots in Figure 3. Toy data set one the species are indistinct using combinations of northing and length. Toy data two, the species can be separated by a linear combination of northing and length and they are present in equal amounts. Toy data three, the species can be separated by northing alone. Toy data four, the species are unable to be separated by a linear combination of northing and length, but the species are separate. One of the species in toy data four is in two distinct groups. The apparent error rate of each method for each toy data set is provided for the model, and the cross validation of the model in Table 23 at the end of this section for model performance comparisons.

Table 4: Variance-covariance matrices for toy data sets for all data, and for each species.

Data set		Pooled		Īnanga		Giant Kōkopu	
	Variable	northing	length	northing	length	northing	length
1	northing	1.000	-0.766	1.082	-0.852	0.769	-0.533
	length	-0.766	1.000	-0.852	1.060	-0.533	0.851
2	northing	1.000	-0.003	0.533	-0.452	0.490	-0.427
	length	-0.003	1.000	-0.452	0.662	-0.427	0.601
3	northing	1.000	-0.766	0.211	-0.170	0.380	-0.388
	length	-0.766	1.000	-0.170	0.401	-0.388	0.926
4	northing	1.000	0.168	0.360	-0.393	1.410	0.473
	length	0.168	1.000	-0.393	0.597	0.473	1.126

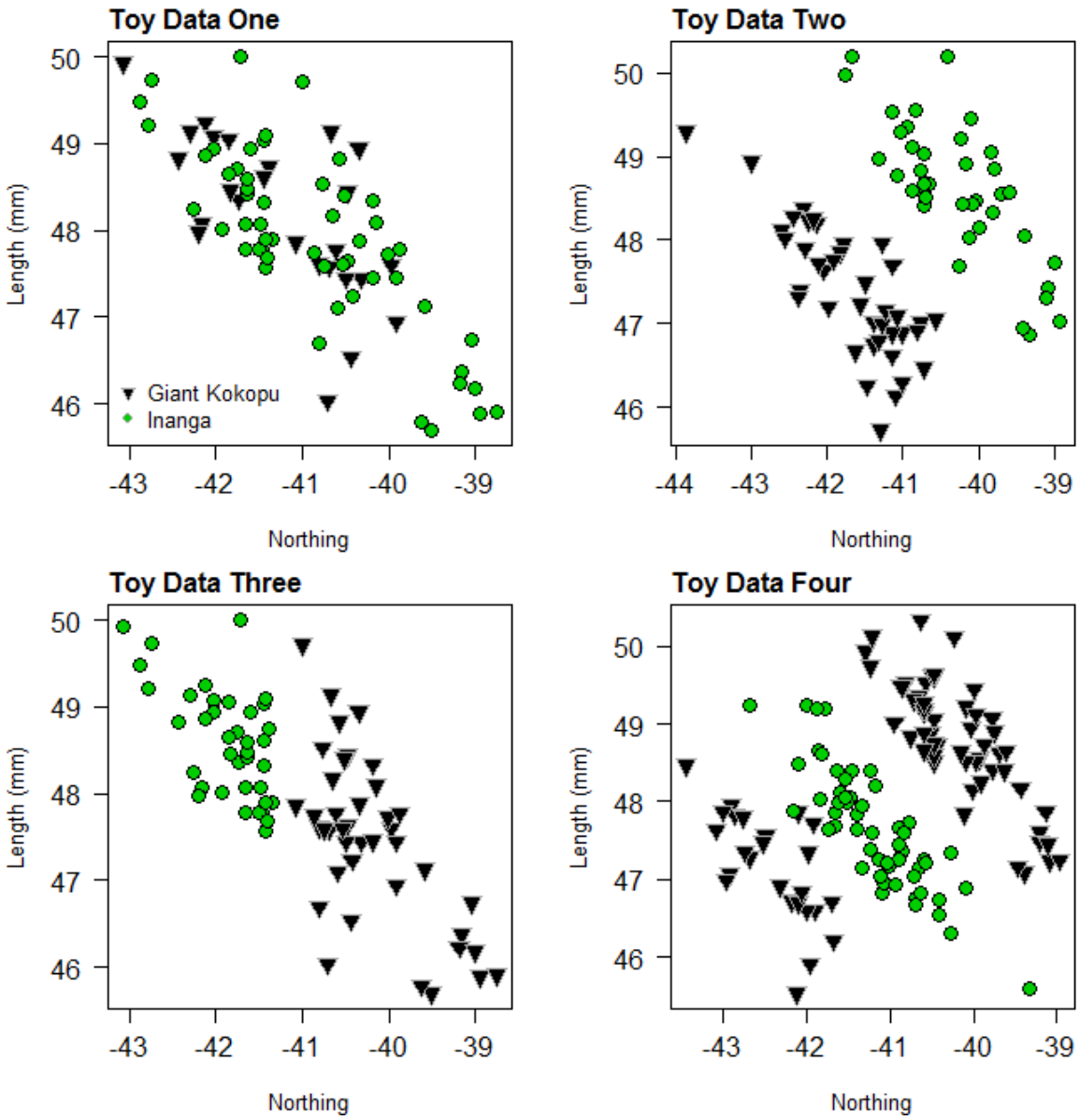


Figure 3: Toy data sets. Toy data set one; the species are not linearly separable. Toy data set two; species are linearly separable by a linear combination of both predictor variables. Toy data three; species are linearly separated by northing. Toy data set four; has high within species variance for one species and low within class variance for the other species.

Multinomial Logistic regression

The result of multinomial logistic regression is a set of probabilities, one for each species. This method takes in the measurements associated with an observation and returns the species the observation is most likely to belong to (Agresti, 2013).

The probability for the Jth species is calculated using Equation 1.

$$\pi_{j=J}(x) = \frac{1}{1+\sum_{c=1}^{J-1} exp(\alpha_c + \beta_c^T x)} \quad [1]$$

And for all other species, Equation 2:

$$\pi_j(x) = \frac{exp(\alpha_c + \beta_c^T x)}{1+\sum_{c=1}^{J-1} exp(\alpha_c + \beta_c^T x)} \quad [2]$$

where π_j is the response probability for class j, x is a vector of all the predictors, α is an intercept term, and β is a slope term. When only two classes are involved, the multinomial logistic regression becomes binary logistic regression, as per Equation 3:

$$pr(c = 1) = 1 - pr(c = 0) = \frac{exp(\alpha+\beta^T x)}{1+ exp(\alpha+\beta^T x)} \quad [3]$$

Examples of Multinomial logistic regression

Fitting logistic regression to the toy data four, produced the coefficients $\alpha = 22.20$ and $\beta = (-0.66, -0.22)$. Using these we can, for example, estimate the probability that a specimen caught at northing -41.49 of length 48.07mm is īnanga as per Equation 4.

$$pr(\bar{\imath}nanga) = \frac{exp(\alpha + \beta^T x)}{1 + exp(\alpha + \beta^T x)}$$

$$= \frac{\exp(22.37+(-0.66,-0.21)^T[-41.5,48.1])}{1+\exp(22.37+(-0.66,-0.21)^T[-41.5,48.1])} = 0.4088 \text{ (4dp)} \quad [4]$$

Each observation is typically allocated to the class with the highest probability multinomial logistic regression. In the toy examples, we have only two species, so we revert to binary logistic regression. Figure 4 shows the results for all toy data sets. Blue areas are where the model would classify observations as īnanga, and the cream areas are where the model would classify observations as giant kōkopu

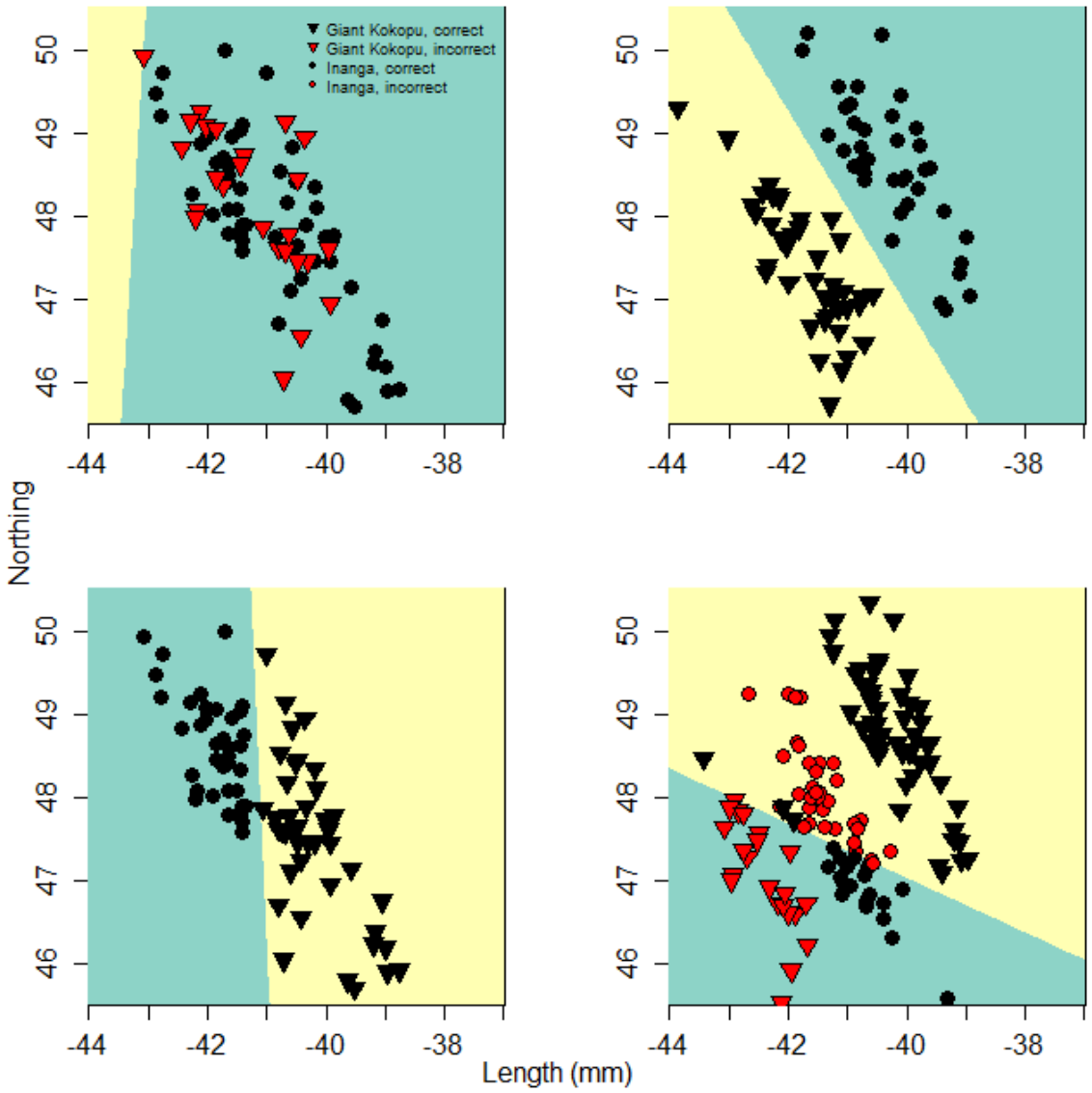


Figure 4: Plots of classification accuracy for MLR on all toy datasets. Species in toy data one were unable to be separated by MLR, so all observations were classified as giant kōkopu. Species in toy data two and toy data three were linearly separable so there were no observations misclassified. The species in toy data four are separate from each other but unable to be distinguished by MLR. All misclassified observations are indicated in red.

Table 5: Confusion matrices for Multinomial Logistic Regression (MLR) on all toy datasets. Species in toy data one were unable to be linearly separated resulting in all observations classified as giant kōkopu. Species in toy data two were linearly separated and each observation was correctly classified, even when cross validated. Species in toy data three were also linearly separable so there were no observations misclassified, even when cross validated. The species in toy data four are not linearly separable so are difficult to discriminate using MLR. The AER was the same as the CV error for all data sets except data set one.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV Error
toyData1			31.64%	34.17%
Predicted Giant Kōkopu	0	0		
Predicted Īnanga	25	54		
toyData2			0.00%	0.00%
Predicted Giant Kōkopu	40	0		
Predicted Īnanga	0	40		
toyData3			0.00%	0.00%
Predicted Giant Kōkopu	41	0		
Predicted Īnanga	0	38		
toyData4			42.11%	42.11%
Predicted Giant Kōkopu	57	34		
Predicted Īnanga	22	20		

Assumptions, Pros and Cons

- Assumes the probabilities are linear in $\boldsymbol{x}$
- No covariates are independent.

Pros:

- Computationally inexpensive.

Cons:

- Does not classify well if the classes are not linearly separable, for example toy data one, toy data four (Table 5).

Linear discriminant analysis (LDA)

Let X be an $n \times k$ matrix of inputs where n is the number of observations and k is the number of covariates, and let Y be the categorical response, putting each observation in one of the C classes. The result of LDA is a set of C linear functions is shown in Equation 4:

$$\delta_C = x^T \Sigma^{-1} \mu_C - \frac{1}{2} \mu_C^T \Sigma^{-1} \mu_C + log\pi_C \quad [4]$$

where x a is a vector of covariates for an individual observation, Σ is the pooled variance-covariance matrix, μ is the vector of means for each variable, and π is $\frac{N_C}{N}$ where N_c is the number of class C observations and N is the total number of observations. The observation is then classified to the class for which the associated function gets the highest value.

Examples of LDA

Consider an example from toy data three. The linear discriminant functions in this case are Eq 2 and Eq 3. If we have an observation with -41.49 northing and length 48.07mm, the function value will be equal to1180.25 and 1181.03 for giant kōkopu and īnanga respectively.

$$\delta_{giantKokopu}(x) = x^T \Sigma^{-1} \mu_{gK} - \frac{1}{2} \mu_{gK}^T \Sigma^{-1} \mu_{gK} + log\pi_{gK} \quad [5]$$

$$= [-41.49, 48.07]^T \begin{bmatrix} 1.00 & -0.77 \\ -0.77 & 1.00 \end{bmatrix} [-40.20, 47.44] - \frac{1}{2} [-40.20, 47.44]^T \begin{bmatrix} 1.00 & -0.77 \\ -0.77 & 1.00 \end{bmatrix} [-40.20, 47.44] - 0.655$$

$$= [-41.49, 48.07]^T [-9.35, -40.29] - 1144.34 = 1180.25$$

$$\delta_{inanga}(x) = x^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + log\pi_i \quad [6]$$

$$= [-41.49, 48.07]^T \begin{bmatrix} 1.00 & -0.77 \\ -0.77 & 1.00 \end{bmatrix} [-41.86, 48.59] - \frac{1}{2} [-41.86, 48.59]^T \begin{bmatrix} 1.00 & -0.77 \\ -0.77 & 1.00 \end{bmatrix} [-41.86, 48.59] - 0.732$$

$$= [-41.49, 48.07]^T [-11.26, 39.96] - 1207.81 = 1181.03$$

Because $\delta_{giantKokopu}(x) < \delta_{inanga}(x)$, the observation will be classified as inānga

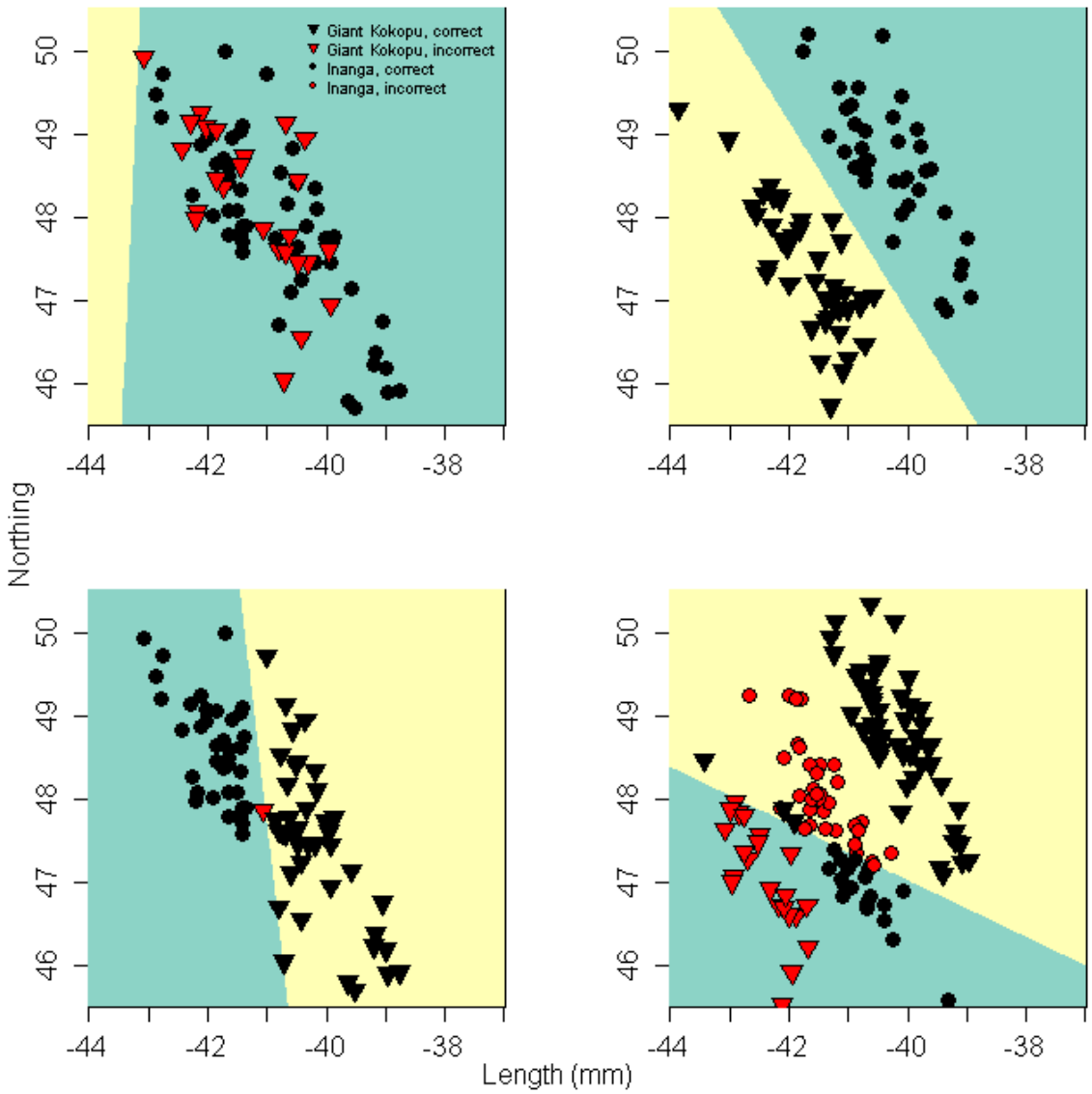


Figure 5: Plots of classification accuracy for LDA on all toy datasets. Classes in toy data one were very mixed as classes were unable to be linearly separated. As result LDA classified all observations as giant kōkopu. Classes in toy data two were linearly separated and each observation was correctly classified. Classes in toy data three were also linearly separable so there was only one observation misclassified. The classes in toy data four are separate from each other but unable to be distinguished by a linear combination of Length and Northing.

Telling the classes apart is easier when the classes are further apart, or when there is less variability within the classes, or when the classes are far apart and have low within-class variability (Agresti, 2003). The results of LDA for the four toy examples in Figure 11 demonstrate this. *Table 6* shows the confusion matrices.

Table 6: Confusion matrices for LDA on all toy datasets. Classes in toy data one were mixed. Classes were unable to be linearly separated. As result LDA classified all observations as giant kōkopu. Classes in toy data two were linearly separated and each observation was correctly classified. Classes in toy data 3 were also linearly separable so there was only one observation misclassified. The classes in toy data 4 are not linearly separable so are difficult to discriminate using LDA. The AER for cross validated models was always equal to, or greater than the AER of the descriptive model.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV Error
toyData1			31.64%	34.18%
Predicted Giant Kōkopu	0	0		
Predicted Īnanga	25	54		
toyData2			0.00%	0.00%
Predicted Giant Kōkopu	40	0		
Predicted Īnanga	0	40		
toyData3			1.27%	2.53%
Predicted Giant Kōkopu	40	0		
Predicted Īnanga	1	38		
toyData4			42.11%	42.86%
Predicted Giant Kōkopu	57	34		
Predicted Īnanga	22	20		

Effect of Species Prevalence in LDA

Sometimes the sample prevalence of species do not reflect population prevalence. Population prevalence may be substituted as $p_i = \widehat{p_i}$ instead. Where $\widehat{p_i}$ is the estimated proportion of species i.

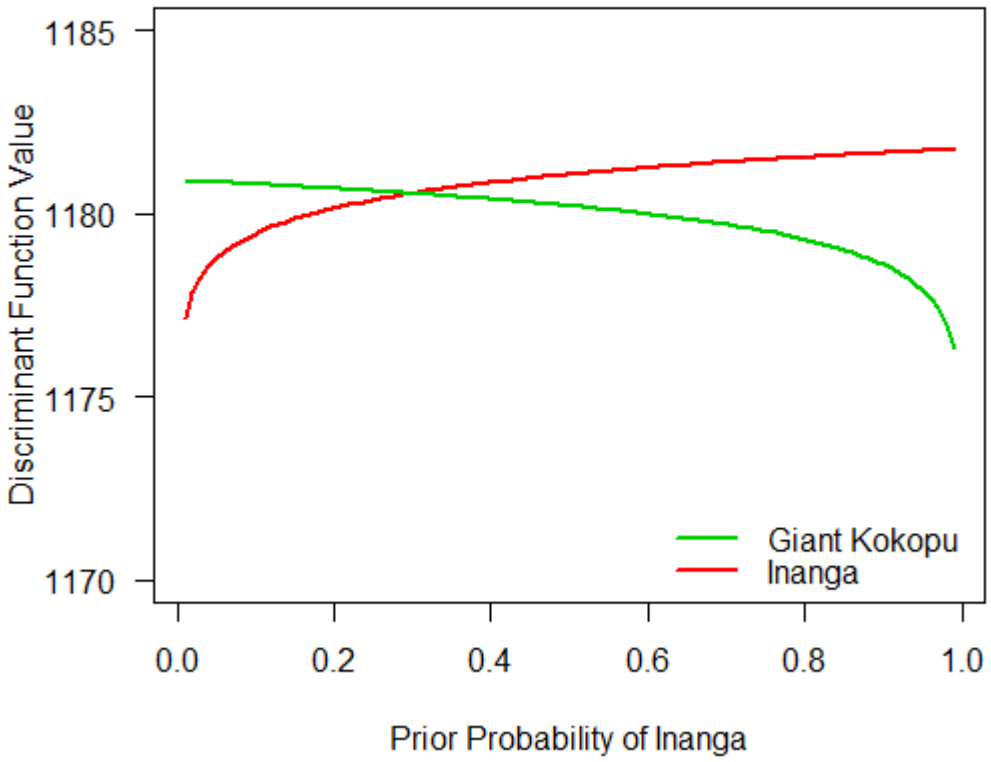


Figure 6: Plotting the discriminant function values when changing the prior probabilities of īnanga for an observation with northing -41.49, and length 48.07mm. As the probability of īnanga increases, it is more likely that the observation will be classified to species īnanga using LDA.

Table 7: Confusion matrix for LDA on all toy data three with altered priors. There are more observations misclassified as īnanga when the prior favours īnanga. More observations are classified as giant kōkopu when the prior is in favour of giant kōkopu. The CV error is less than the AER for the model that favours giant kōkopu.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV Error
toyData2 – Probability: Giant Kōkopu = 0.99, Īnanga = 0.01			27.85%	26.58%
Predicted Giant Kōkopu	41	22		
Predicted Īnanga	0	16		
toyData2 – Probability: Giant Kōkopu = 0.01, Īnanga = 0.99			29.11%	29.11%
Predicted Giant Kōkopu	18	0		
Predicted Īnanga	23	38		

Changing the prior probability of each class Changes the position of the cut off between class predictions for the LDA model (Figure 7).

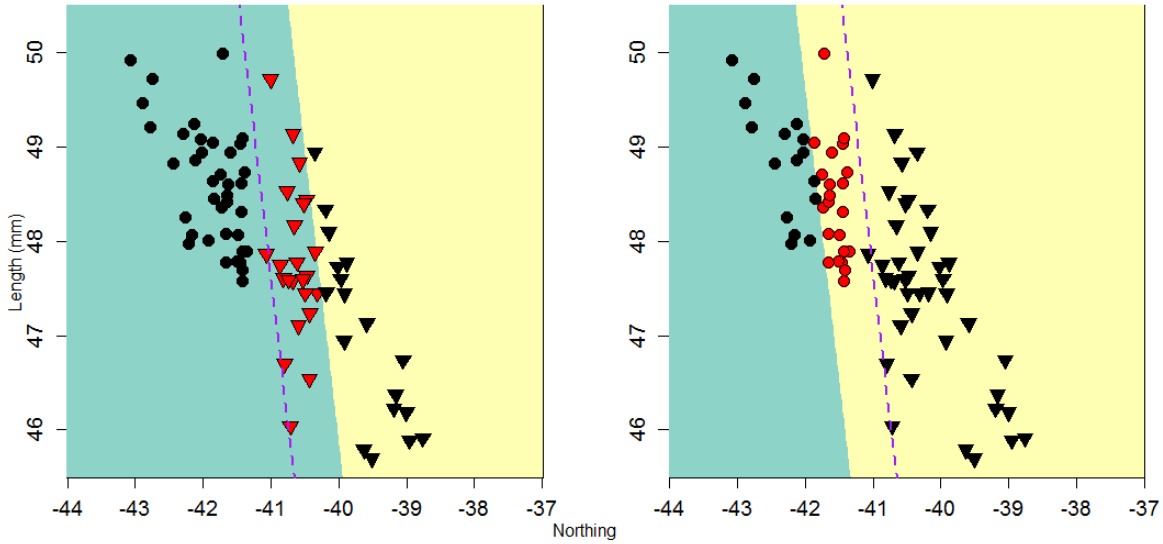


Figure 7: Toy data set three with changed prior probabilities of each class. Panel A demonstrates how classifications change when the priors are in favour of giant kōkopu, there are more observations classified as giant kōkopu. In panel A demonstrates how classifications change when the priors are in favour of īnanga, there are more observations classified as īnanga. The dotted purple lines show where LDA would discriminate with natural sample proportions of each species.

Assumptions, Pros and Cons

- The covariates have a common (not class-specific) variance-covariance matrix,
- Predictors come from a multivariate normal distribution

Pros:

- Computationally inexpensive.

Cons

- Cannot use categorical predictors as categorical predictors do not come from a multi-variate normal distribution.
- Does not classify well if the classes are not linearly separable – for example toy data one, toy data four (Figure 5 and Table 6).

Quadratic discriminant analysis (QDA)

Let X be an $n \times k$ matrix of inputs where n is the number of observations and k is the number of covariates, and let Y be the categorical response, putting each observation in one of the C classes. To calculate the discriminant function for each class we would use Equation 7.

$$\delta_c(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c) + log\pi_c \quad [7]$$

where x a is a vector of covariates for an individual observation, Σ_c is the variance-covariance matrix of the class, μ is the vector of means for each variable, and π is $\frac{N_C}{N}$ where N_c is the number of class C observations and N is the total number of observations. The observation is then classified to the class for which the associated function gets the highest value.

Examples of QDA

Consider an example from toy data three. The quadratic discriminant functions in this case are Equation 8 and Equation 9. If we have an observation with -41.49 northing and length 48.07mm, the function value will be equal to -2.704 and -3.791 for giant kōkopu and īnanga respectively.

$$\delta_{giantKokopu}(x) = -\frac{1}{2}log|\Sigma_{gK}| - \frac{1}{2}(x-\mu_{gK})^T \Sigma_{gK}^{-1}(x-\mu_{gK}) + log\pi_{gK} \quad [8]$$

$$= 1.09 - \frac{1}{2}[0.199, 1.009]^T \begin{bmatrix} 5.37 & 3.82 \\ 3.82 & 4.38 \end{bmatrix} [0.199, 1.009] + log(0.5)$$

$$= 1.09 - 3.11 - 0.69 = -2.704\ (3dp)$$

$$\delta_{giantKokopu}(x) = -\frac{1}{2}\log|\Sigma_{gK}| - \frac{1}{2}(x-\mu_{gK})^T \Sigma_{gK}^{-1}(x-\mu_{gK}) + log\pi_{gK} \quad [9]$$

$$= 0.95 - \frac{1}{2}[-1.20, -0.21]^T \begin{bmatrix} 4.46 & 3.05 \\ 3.05 & 3.59 \end{bmatrix} [-1.20, -0.21] + log(0.5)$$

$$= 0.95 - 4.05 - (-0.69) = -3.791\ (3dp)$$

Because $\delta_{giantKokopu}(x) > \delta_{inanga}(x)$ the observation will be classified as giant kōkopu.

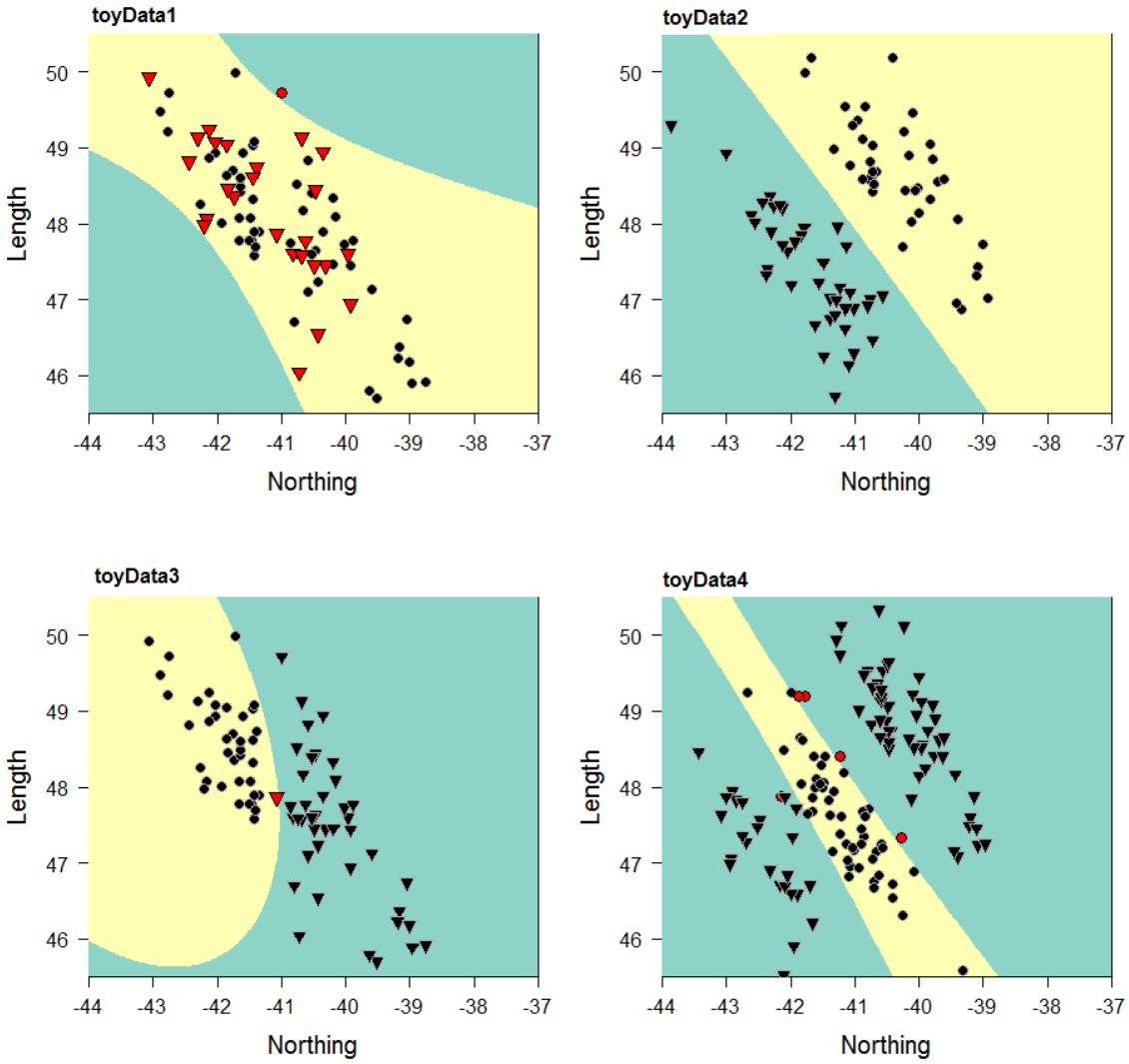


Figure 8: Plots of classification accuracy for QDA on all toy datasets. Species in toy data one were unable to be linearly separated. As result QDA classified all observations as giant kōkopu. Classes in toy data two were linearly separable by a combination of Northing and Length. Each observation was correctly classified. Classes in toy data three were linearly separable by Northing so there was only one observation misclassified. The species in toy data four are separate from each other but unable to be distinguished by a linear combination of Length and Northing.

Telling the classes apart is easier when the classes are further apart, when there is less variability within the classes, or when the classes are far apart and have low within-class variability (Agresti, 2003). The results of QDA for the four toy examples in Figure 8 demonstrate this. Table 8 shows the confusion matrices. The species in toy data one are not separate and were not classified well with QDA, the AER and CV error are over 30%. Toy data two has the lowest AER and CV error. The species are separate. The species are separated by a linear combination of Northing and Length in toy data three. The AER and CV error for toy data there is less than 3%. Toy data four, the species are separate but giant kōkopu has high variance. The AER was 3.76% and the CV error was 6.01% (Table 8).

Table 8: Confusion matrices for QDA on all toy datasets. Classes in toy data one were mixed. Classes were unable to be linearly separated. As result QDA classified all observations as giant kōkopu. Classes in toy data two were linearly separated and each observation was correctly classified. Classes in toy data three were also linearly separable so there was only one observation misclassified. The classes in toy data four are not linearly separable so are difficult to discriminate using QDA. The CV error was always equal to, or greater than the AER.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV error
toyData1			32.91%	35.44%
Predicted Giant Kōkopu	0	1		
Predicted Īnanga	25	53		
toyData2			0.00%	0.00%
Predicted Giant Kōkopu	40	0		
Predicted Īnanga	0	40		
toyData3			1.27%	2.53%
Predicted Giant Kōkopu	40	0		
Predicted Īnanga	1	38		
toyData4			3.76%	6.01%
Predicted Giant Kōkopu	79	5		
Predicted Īnanga	0	49		

Effect of Species Prevalence in QDA

Sometimes the frequencies in the sample do not reflect population prevalence. Sample prevalence may be substituted as $p_i = \widehat{p_i}$ instead. Where $\widehat{p_i}$ is the estimated proportion of species i. There are more observations misclassified as īnanga when the prior favours īnanga (Table 9) (Figure 9). More observations are classified as giant kōkopu when the prior is in favour of giant kōkopu. The CV error is less than the AER for the model that favours giant kōkopu (Table 9).

Changing the prevalence of each class changes the position of the cut off between class predictions for the QDA model (Figure 9). For the observation with a northing -41.49, and length 48.07mm from

toy data two, as the prevalence of īnanga increases, it is more likely that will be classified as īnanga using QDA as the value of the discriminant function for īnanga gets larger, and the discriminant function for giant kōkopu gets smaller.

Table 9: Confusion matrix for QDA on all toy data two with altered priors. There are more observations misclassified as īnanga when the prior favours īnanga. No observations are misclassified as giant kōkopu when the prior is in favour of giant kōkopu.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV Error
toyData2 – Proportions: Giant Kōkopu = 0.99, Īnanga = 0.01				
Predicted Giant Kōkopu	40	0	0.00%	0.00%
Predicted Īnanga	0	40		
toyData2 – Proportions: Giant Kōkopu = 0.01, Īnanga = 0.99				
Predicted Īnanga	37	0	3.75%	3.75%
Predicted Giant Kōkopu	3	40		

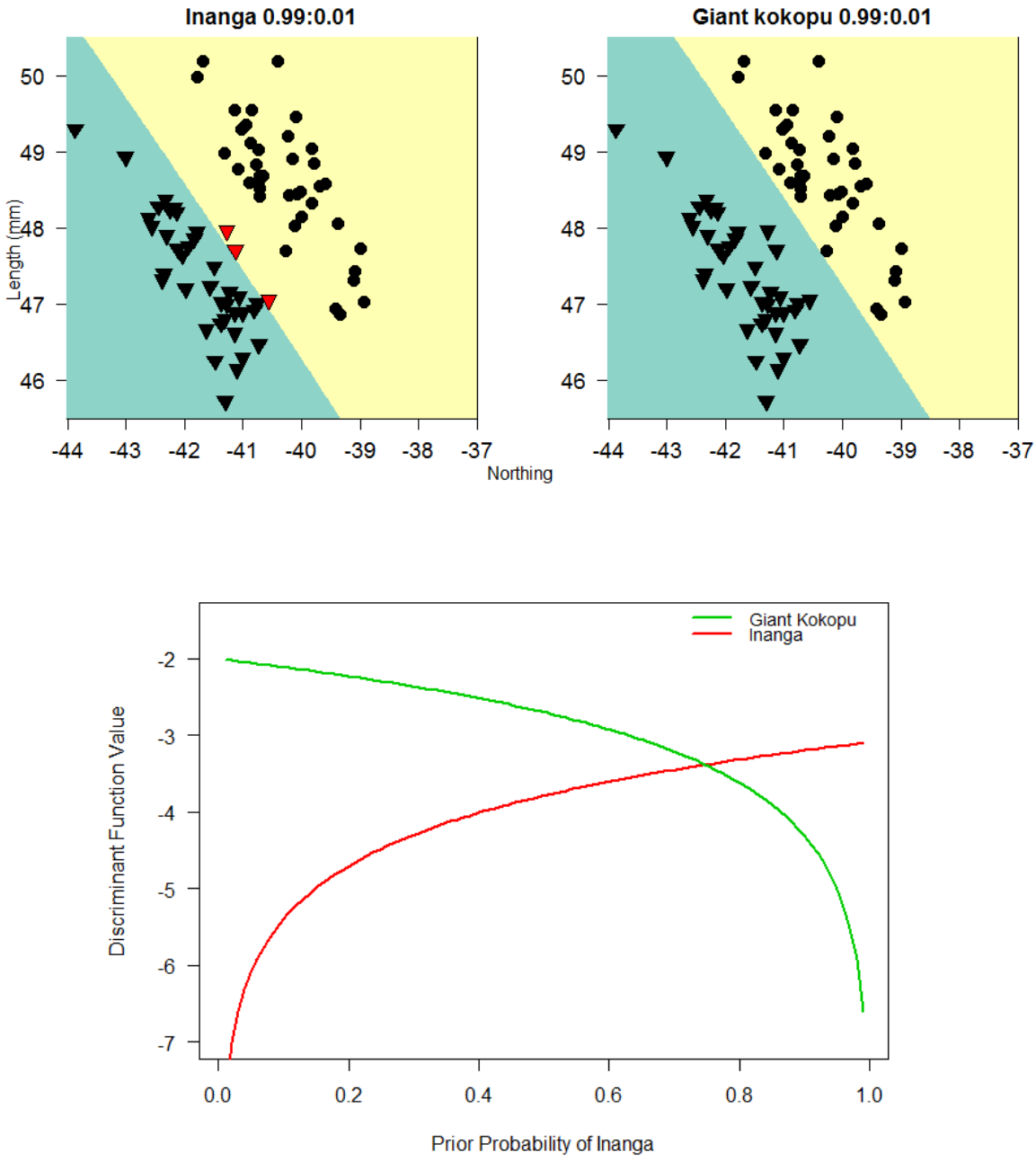


Figure 9: Plotting the changing discriminant function values for changing the prevalence of īnanga for an observation of northing -41.49, and length 48.07mm from toy data two. The top left panel demonstrates how classifications change when the priors are in favour of īnanga, three giant kōkopu are misclassified as īnanga. The right panel demonstrates how classifications change when the priors are in favour of giant kōkopu. No observations are misclassified. The bottom panel shows that as the prevalence of īnanga increases, it is more likely that will be classified as īnanga using QDA as the value of the discriminant function for īnanga gets larger, and the discriminant function for giant kōkopu gets smaller.

Assumptions, Pros and Cons

- The covariates have a class-specific variance-covariance matrix
- No multicollinearity of covariates Invertible covariance matrices
- Predictors come from a multivariate normal distribution

Pros:

- Computationally inexpensive.

Cons

- Cannot use categorical predictors as categorical predictors do not come from a multi-variate normal distribution.
- Does not classify well if the classes are not separable – for example toy data one (Figure 8 and Table 8).

Naïve Bayes

Let X be an $n \times k$ matrix of inputs where n is the number of observations and k is the number of covariates, and let Y be the categorical response, putting each observation in one of the C classes.

We assume, that the covariates x are from a multivariate normal distribution, $X \sim N(\mu, \Sigma)$ where μ is a vector of means for all covariates, and Σ is the variance-covariance matrix for all covariates of one of the species.

The posterior probability of an observation being classified to a species can be calculated using Bayes Rule (Equation 10).

$$pr(Y|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|Y)pr(Y)}{\sum_{c=1}^{C} f(\boldsymbol{x}|Y_c)pr(Y_c)}$$

$$= \frac{\left(\frac{1}{2\pi|\boldsymbol{\Sigma_1}|^{\frac{1}{2}}} e^{-(x-\boldsymbol{\mu_1})'\boldsymbol{\Sigma_1}^{-1}(x-\boldsymbol{\mu_1})}\right) * pr(y=c_1)}{\left(\frac{1}{2\pi|\boldsymbol{\Sigma_1}|^{\frac{1}{2}}} e^{-(x-\boldsymbol{\mu_1})'\boldsymbol{\Sigma_1}^{-1}(x-\boldsymbol{\mu_1})}\right) * pr(y=c_1) + \left(\frac{1}{2\pi|\boldsymbol{\Sigma_2}|^{\frac{1}{2}}} e^{-(x-\boldsymbol{\mu_2})'\boldsymbol{\Sigma_2}^{-1}(x-\boldsymbol{\mu_2})}\right) * pr(y=c_2)} \quad [10]$$

Examples of Naïve Bayes

For example, given an observation has northing -41.5 and length 47mm from toy data four, the posterior probability of it being īnanga or giant kōkopu can be calculated as per Equations 11 and 12

$$pr(\textit{inānga}\,|{-41.5}\ \textit{north}, 47mm) = \frac{pr(-41.5, 47|\textit{inānga})pr(\textit{inānga})}{\Sigma pr(-41.5\,, 47|\textit{inānga})pr(\textit{inānga}), (-41.5\,, 47|\textit{giant kōkopu})pr(\textit{giant kōkopu})}$$

$$= \frac{\left(\frac{1}{2\pi|\boldsymbol{\Sigma_1}|^{\frac{1}{2}}} e^{-\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_1}\right)'\boldsymbol{\Sigma_1}^{-1}\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_1}\right)}\right) * \frac{54}{133}}{\left(\frac{1}{2\pi|\boldsymbol{\Sigma_1}|^{\frac{1}{2}}} e^{-\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_1}\right)'\boldsymbol{\Sigma_1}^{-1}\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_1}\right)}\right) * \frac{54}{133} + \left(\frac{1}{2\pi|\boldsymbol{\Sigma_2}|^{\frac{1}{2}}} e^{-\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_2}\right)'\boldsymbol{\Sigma_2}^{-1}\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_2}\right)}\right) * \frac{79}{133}} \quad [11]$$

$$= 0.247\ (3dp)$$

$$pr(\textit{giant kōkopu}\,|{-41.5}\ \textit{north}, 47mm) = \frac{pr(-41.5, 47|\textit{giant kōkopu})pr(\textit{giant kōkopu})}{\Sigma pr(-41.5\,, 47|\textit{inānga})pr(\textit{inānga}), (-41.5\,, 47|\textit{giant kōkopu})pr(\textit{giant kōkopu})}$$

$$= \frac{\left(\frac{1}{2\pi|\boldsymbol{\Sigma_2}|^{\frac{1}{2}}} e^{-\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_2}\right)' \boldsymbol{\Sigma_2}^{-1}\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_2}\right)}\right) * \frac{79}{133}}{\left(\frac{1}{2\pi|\boldsymbol{\Sigma_1}|^{\frac{1}{2}}} e^{-\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_1}\right)' \boldsymbol{\Sigma_1}^{-1}\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_1}\right)}\right) * \frac{54}{133} + \left(\frac{1}{2\pi|\boldsymbol{\Sigma_2}|^{\frac{1}{2}}} e^{-\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_2}\right)' \boldsymbol{\Sigma_2}^{-1}\left(\binom{-41.5}{47.0}-\boldsymbol{\mu_2}\right)}\right) * \frac{79}{133}} \quad [12]$$

= 0.753 (3dp)

where $\mu_{inānga} = \begin{pmatrix} -41.2 & 47.6 \end{pmatrix}$ and $\Sigma_{inānga} = \begin{bmatrix} 0.36 & -0.39 \\ -0.39 & 0.60 \end{bmatrix}$

and $\mu_{giantkōkopu} = \begin{pmatrix} -40.9 & 48.3 \end{pmatrix}$ and $\Sigma_{giantkōkopu} = \begin{bmatrix} 1.41 & 0.47 \\ 0.47 & 1.13 \end{bmatrix}$

Parameters μ_c and Σ_c which are estimated from sample data (Table 3, and Table 4). Parameters can be also estimated from previous studies. The observation with northing -41.5 and length 47mm from toy data four would be classified as giant kōkopu as the posterior probability of giant kōkopu is greater than the posterior probability of īnanga.

Naïve Bayes extends this idea to produce a set of C probabilitieswhere $pr(x)$ is a function of y and x that is proportional to the probability of an observation belonging to class y. $pr(y = c_i)$ is calculated from a multivariate normal distribution with parameters μ_c and Σ_c which are estimated from sample data, or from prior studies. This is known as Gaussian Naïve Bayes.

Figure 10 demonstrates how Naïve Bayes would perform with the toy data sets. Species in toy data one were very mixed as species were unable to be separated. As result Naïve Bayes classified all observations as giant kōkopu (Table 10). Species in toy data two were linearly separated and each observation was correctly classified and CV error was 0% (Table 10). Species in toy data three were linearly separable by Northing and one observation was misclassified, CV error was 1.27% (Table 10). The species in toy data four are separate from each other but Naïve Bayes CV error was 18.8% (Table 10)

Table 10: Confusion matrices for Naïve Bayes on all toy data sets. Species in toy data one were mixed. Naïve Bayes classified all observations as giant kōkopu. Species in toy data two were linearly separated by Northing and each observation was correctly classified. Species in toy data three were linearly separable. There was one giant kōkopu misclassified as īnanga. The species in toy data four are not linearly separable but were more easily distinguished using Naïve Bayes. The CV error was always equal to, or greater than AER. The CV error was the same as AER where the two species were linearly separable.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV Error
toyData1			31.65%	36.71%
Predicted Giant Kōkopu	0	0		
Predicted Īnanga	25	54		
toyData2			0.00%	0.00%
Predicted Giant Kōkopu	40	0		
Predicted Īnanga	0	40		
toyData3			1.27%	1.27%
Predicted Giant Kōkopu	40	0		
Predicted Īnanga	1	38		
toyData4			17.29%	18.80%
Predicted Giant Kōkopu	68	12		
Predicted Īnanga	11	42		

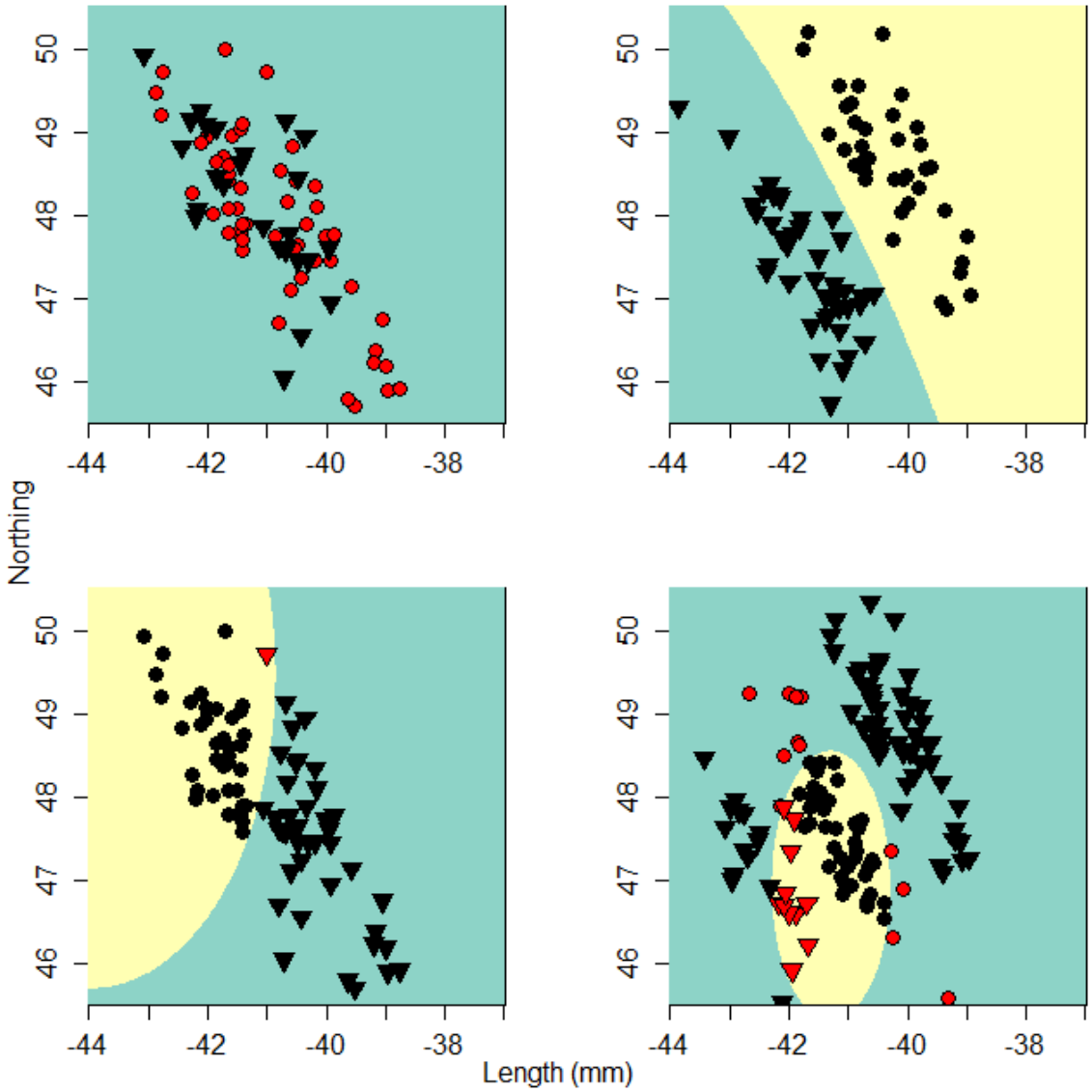


Figure 10: Plots of classification accuracy for Naïve Bayes on all toy datasets. Species in toy data one were very mixed as species were unable to be separated. As result Naïve Bayes classified all observations as giant kōkopu. Species in toy data two were linearly separated and each observation was correctly classified. Species in toy data three were linearly separable by Northing and one observation was misclassified. The species in toy data four are separate from each other and are more readily distinguished by Naïve Bayes than by LDA or MLR.

Discretisation of Continuous Variables

If the distribution of x cannot be assumed to be normal, we can discretise. This means that we divide values of x into categories and calculate observed frequencies instead of using the normal probability distribution f as shown in Figure 11.

We can then use the Bayes Rule for discrete distributions:

$$pr(Y|\boldsymbol{x}) = \frac{pr(\boldsymbol{x}|Y)pr(Y)}{\sum_{c=1}^{C} pr(\boldsymbol{x}|Y_c)pr(Y_c)} \quad [13]$$

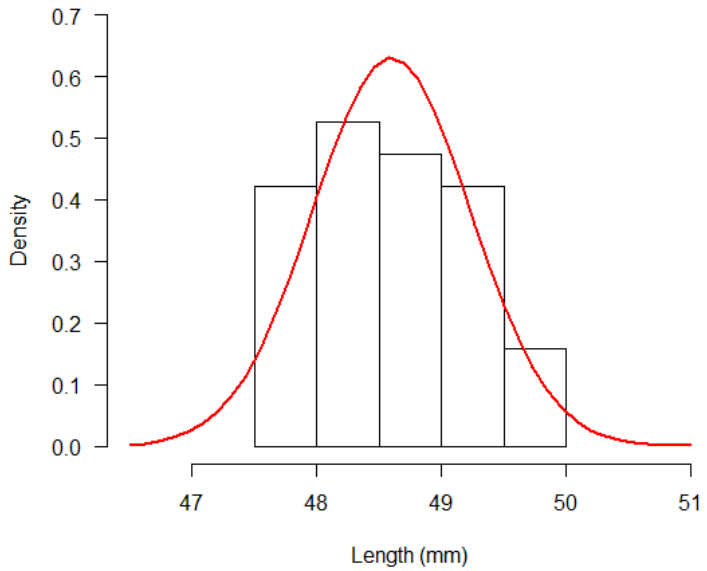


Figure 11: A histogram of length from toy data set four. Normal density curve has been applied over the histogram to show how data with the same parameters would look if it were normally distributed. Naïve Bayes may not perform at its best with data that is not approximately normally distributed.

Table 11 shows how the posterior probabilities change as the resolution of the discretisation changes. A wider resolution increases the chances of representing a species that is dissimilar from other species that are in the same discretised class. To deal with empty categories, a small number, or smoothing factor (δ), is often added to all the observed frequencies. A smoothing factor adds a pseudo count to each category (Figure 12). This pseudo-count is usually a very small number, usually less than one. As the value of the smoother increases, the affect it has on classification probabilities increases (Table 11). Setting δ close to zero does not alter the posterior probabilities a great deal. However, setting δ as increasingly larger numbers alters the posterior probabilities considerably (Table 11). Naïve Bayes can also accommodate variables which are already discrete such as colour (Hastie & Tibshirani, 2009)

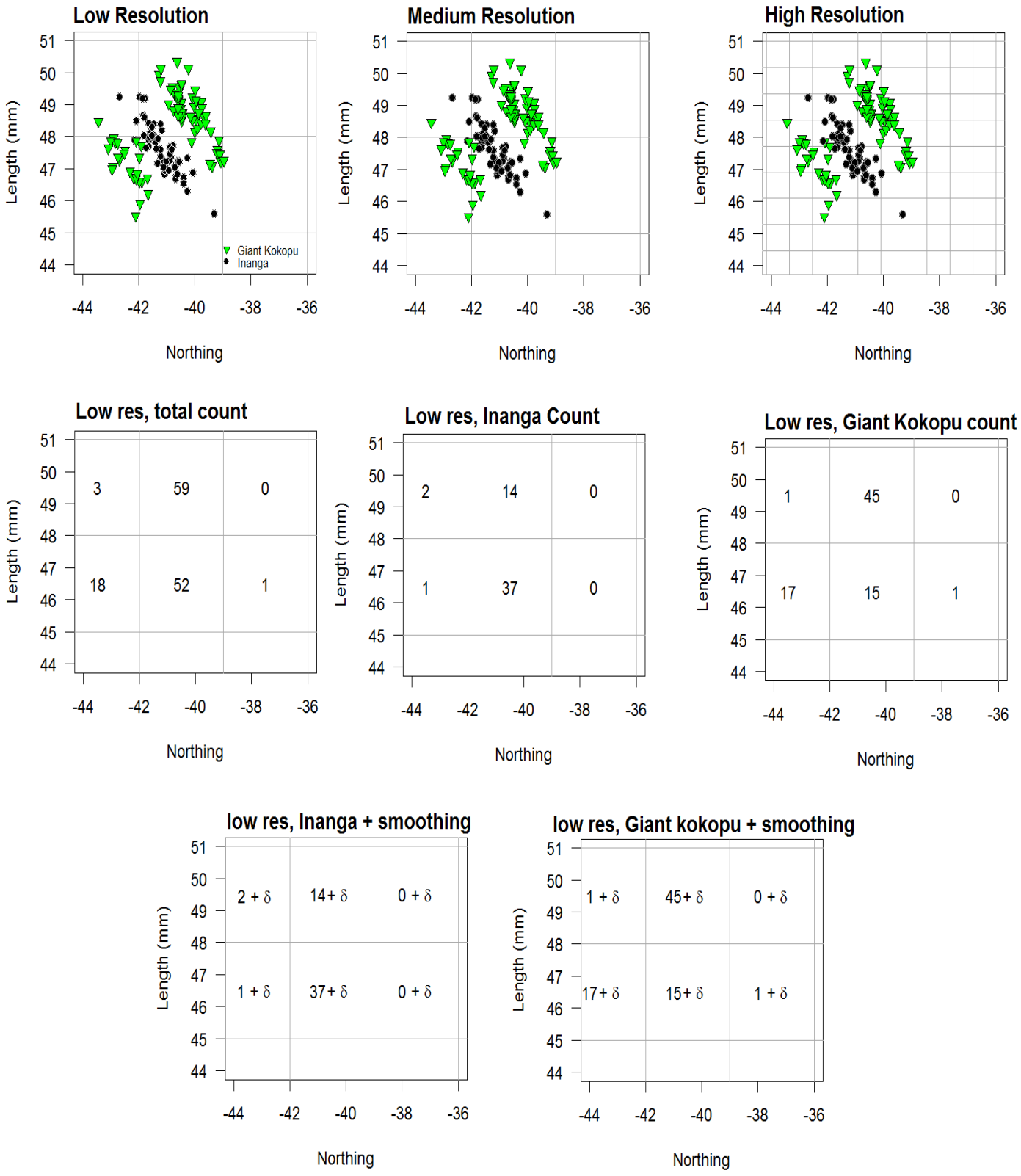


Figure 12: Three different resolutions of discretisation for toy data four and below with counts for low resolution. The left panel shows counts for both species, īnanga in the middle, and giant kōkopu on the right. There are no observations for the class that would contain the measurement Northing-38, and Length 46mm for īnanga. The smoothing as shown in the bottom left panel makes the Naïve Bayes calculation possible.

Table 11: Posterior probabilities of each species for an observation with measurement northing -41.9 and length 47.9mm from toy data four. Posterior probabilities are different dependent upon the resolution, and smoothing factor. With all resolutions for this observation, having a large smoothing factor changes which species has the larger posterior probability, and how the observation would be classified.

Resolution, (Northing x Length classes) Figure 12	Smoothing factor (δ) Figure 12	Posterior probability (4dp)	
		Īnanga	Giant Kōkopu
Low (3 x 2 classes)	none	0.7115	0.2885
	1.0000	0.7115	0.2885
	0.0001	0.4920	0.5081
Medium (6 x 6 classes)	none	0.8750	0.0750
	1.0000	0.8750	0.0750
	0.0001	0.4777	0.4740
High (10 x 9 classes)	none	0.7500	0.2500
	1.0000	0.7489	0.2511
	0.0001	0.4161	0.5839

Effect of Prior Probabilities in Naïve Bayes

Sometimes the sample proportions of the species do not reflect species prevalence. Population frequencies may be considered by changing the prior probabilities. Consider toy data three; if we adjust the prior probabilities to favour īnanga 0.99:0.01 the fish with northing –41.9 and length 48.5mm the observation will be classified as īnanga (Figure 13, Table 12). Changing the prior probabilities to favour giant kōkopu 0.99:0.01 for the fish with northing -41.49, and length 48.07mm, the functions will be equal to 0.071 for giant kōkopu and 0.031 for īnanga. It will therefore be classified to giant kōkopu. Any of the probabilities that are input to the Naïve Bayes formula can be estimated from any source. Favouring either species in extreme measures will make Naïve Bayes classify more observations to the species that has the higher prior probability (Table 12).

Table 12: Confusion matrix for Naive Bayes on all toy data four with altered priors. There are more observations misclassified as īnanga when the prior favours īnanga. More observations are classified as giant kōkopu when the prior is in favour of giant kōkopu.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV Error
toyData4 – Priors: Giant Kōkopu = 0.01, Īnanga = 0.99			40.60%	59.40%
Predicted Giant Kōkopu	79	54		
Predicted Īnanga	0	0		
toyData4 – Priors: Giant Kōkopu = 0.99, Īnanga = 0.01			59.40%	40.60%
Predicted Giant Kōkopu	0	0		
Predicted Īnanga	79	54		

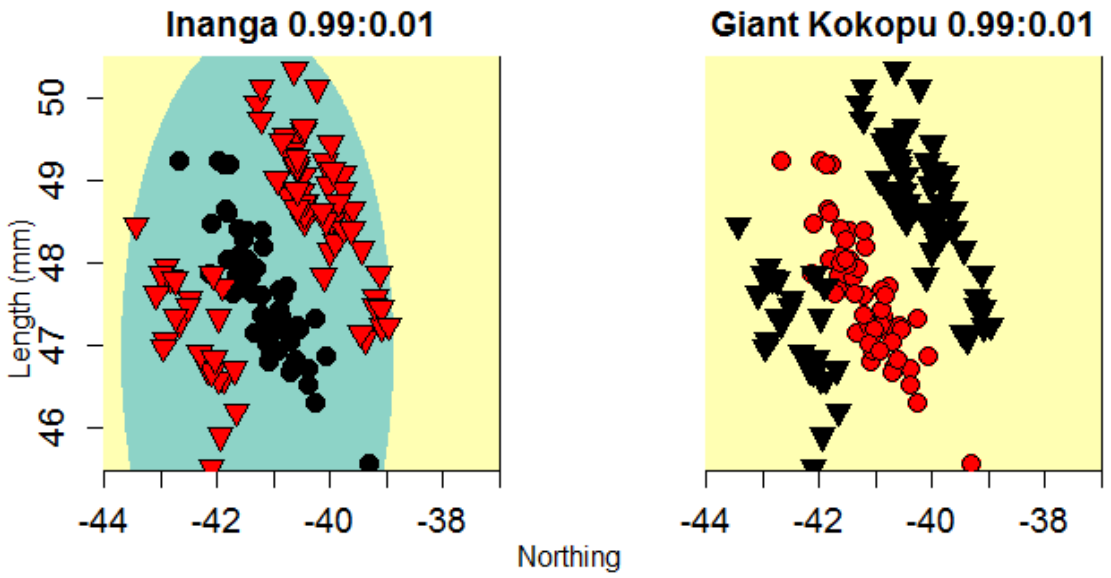


Figure 13: Toy data set four with changed prior probabilities of each class. The left demonstrates how classifications change when the priors are in favour of īnanga. The right demonstrates how classifications change when the priors are in favour of giant kōkopu. There are no correctly classified observations for giant kōkopu when the prior favours īnanga 99%, and there are no correctly classified īnanga when the prior favours giant kōkopu 99%.

Assumptions, Pros and Cons

Assumptions:

For the discrete case there are no assumptions. For continuous predictors, Naïve Bayes can use any appropriate distribution in place of $f(\boldsymbol{x}|Y)$. For Gaussian Naïve Bayes, a multivariate normal distribution is assumed.

Pros

- Naïve Bayes can separate out classes that are not linearly separable. For example, toyData4 has some īnanga observations that are surrounded by giant kōkopu, observations. A single line cannot distinguish between the two classes but Naïve Bayes could pick out that there was high concentration of īnanga in between a group of giant kōkopu
- Continuous, discrete, or categorical variables can be used as predictors
- Prior probabilities can be adjusted to match what the real-life population probabilities are or are estimated to be.

Cons

- Continuous features are commonly turned into discrete variables, but there is no way to determine optimal discretisation as the 'correct level' of discretisation can be subjective.

Higher dimension data (data with lots of predictors) are difficult to classify without a large number of observations. There is a high chance of having empty categories (categories with a frequency or count of observations that are zero) when the distribution has been discretised. Categories with frequencies of zero are not dealt with well within the method (Kotsiantis, 2007) unless a small smoothing factor is specified.

Decision Tree

This method produces a flow chart – like decision tree that directs you to an observation's most likely class using recursive binary splitting (Figure 14). Each node is characterised by the most typical species at that node. Each node in Figure 14 has a number under the species name. This is the purity of that species at that node. Purity is the proportion of the most common species at that node. The bottom number is the percentage of the data that is in that node (Milborrow, 2018).

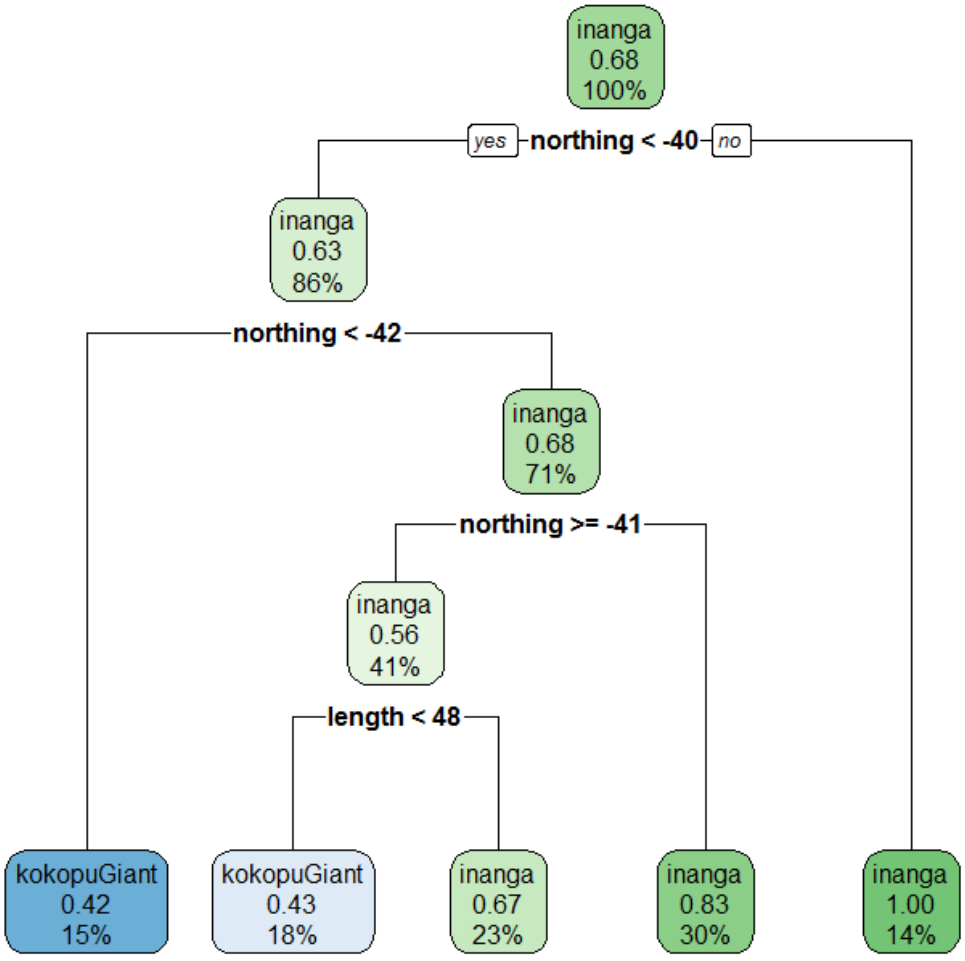


Figure 14: Decision tree with default settings for toy data set one. Each node is a coloured shape. The species at the top of the shape is the most likely species in that node. The next number is the purity of the observations for that species. The bottom number is the percentage of the dataset that is in that node.

The decision tree general algorithm is:

1. Search each covariate for the one that gives the most information. This is the root node and is found using the entropy or information of each predictor variable as calculated by Equation 15 (Friedman, Kohavi, & Yun, 1996).

$$H(Y) = \sum p(y) log\, p(y) \quad [14]$$

   where p is the purity of each class.

2. Place the best covariate at the root of the tree.
3. Split the data into subsets, ideally so that one of the subsets contains just one class for the same value of a predictor variable.
4. Repeat steps 1 and 3 until leaf/terminal nodes are found for the tree.

Information, or entropy, gives the amount of uncertainty of the predictor variable. At each node there can be restrictions placed on the algorithm that limit how a valid split can be made. Each of the nodes contains a logic gate that works towards deciding the class of an observation (James, Witten, Hastie, & Tibshirani, 2013).

Examples of Decision Tree

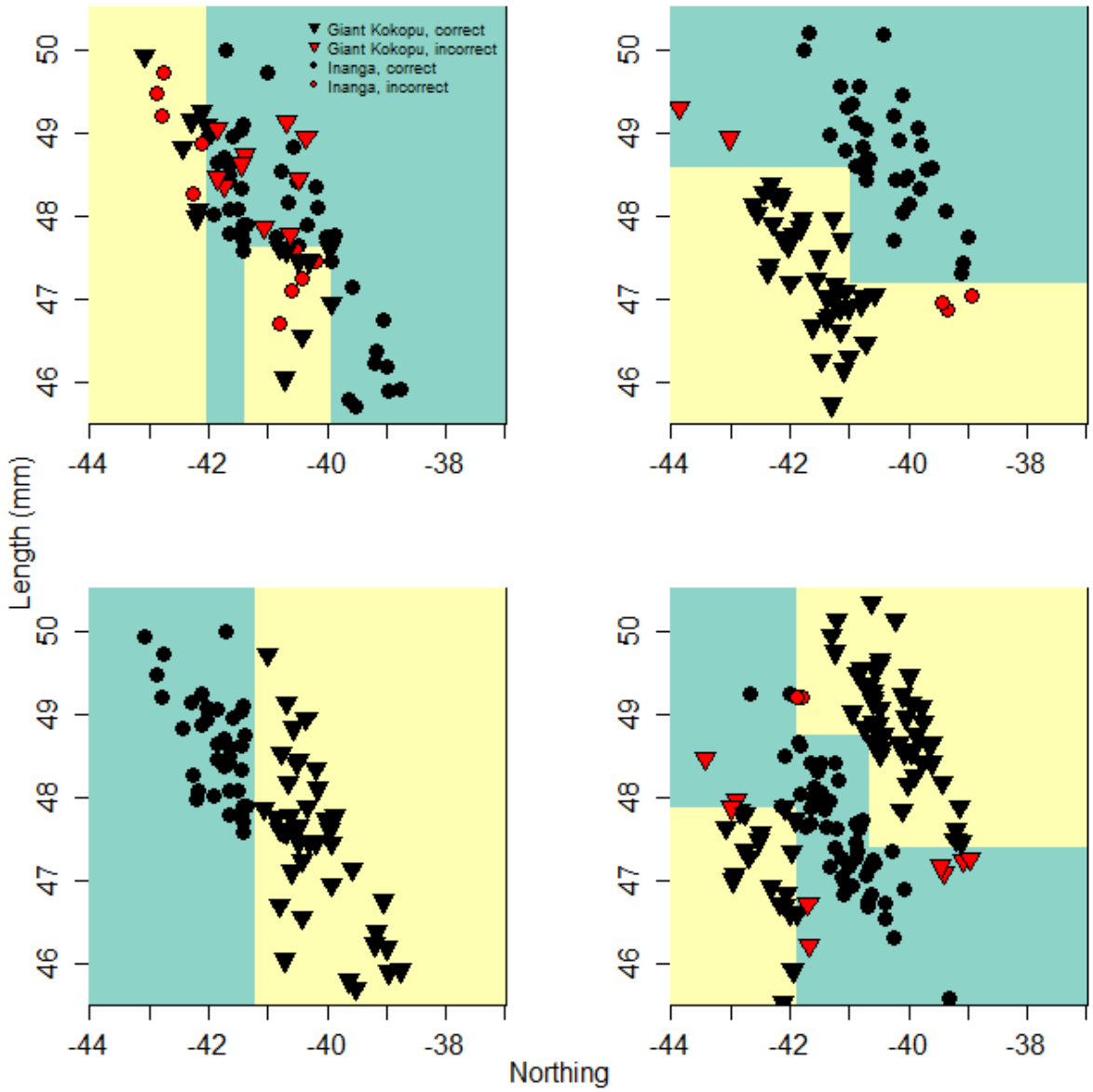


Figure 15: Plots of classification accuracy for decision tree on all toy datasets. Species in toy data one were mixed, species were unable to be linearly separated. As result decision tree misclassified some observations. Species in toy data two were linearly separated, but as a linear combination of two variables. As result decision tree misclassified some observations. Classes in toy data 3 were linearly separable on one variable so there were no misclassifications. The classes in toy data 4 are not linearly separable. The addition of more discrimination rules (branches) increases the classification accuracy but does give 100% perfect classifications.

Let us take an observation from toy data one which has northing of -41.49, and length 48.07mm. If we follow through decision of the tree in Figure 14 for a fish at -41.5 northing and length of 47mm, we see follow this decision path:

1. Northing is less than -40;
2. Northing is not less than -42;

3. Northing is not less than or equal to -41;

Therefore, this observation is classified as īnanga.

Table 13: Confusion matrix for decision tree classifier on all toy datasets. Classes in toy data one were mixed. About a quarter of observations were misclassified. The model for toy data one appears to be over fit as the CV error is considerable larger than the AER. Classes in toy data two were linearly separated but five observations were misclassified. Classes in toy data three were linearly separable. All observations were classified correctly. Giant kōkopu in toy data four is in two clusters but AER and CV error is less than 10%.

	Simulated Giant Kōkopu	Simulated īnanga	AER	CV Error
toyData1			26.58%	45.57%
Predicted Giant Kōkopu	15	11		
Predicted īnanga	10	43		
toyData2			6.25%	13.75%
Predicted Giant Kōkopu	38	3		
Predicted īnanga	2	37		
toyData3			0.00%	0.00%
Predicted Giant Kōkopu	41	0		
Predicted īnanga	0	38		
toyData4			8.27%	9.02%
Predicted Giant Kōkopu	70	2		
Predicted īnanga	9	52		

How decision tree performs on all toy data sets is plotted in Figure 15. Species in toy data one were mixed, species were unable to be linearly separated.. As result decision tree AER is 26.58% and CV error is 45.57% (Table 13). Species in toy data two were linearly separated, but as a linear combination of two variables. AER for toy data two was 6.25 %, and CV error was 45.57%. Classes in toy data three were linearly separable on one variable so there were no misclassifications (Figure 15). The classes in toy data four are not linearly separable. The addition of more discrimination rules (branches) increases the classification accuracy but does give 100% perfect classifications Table 15).

Pruning the Tree

Trees can be grown then pruned. Pruning means that you cut off branches that are unnecessary to the final model. Understanding pruning helps to balance accuracy versus complexity. Growing a tree that is overly complex will describe labelled data well but may not be capable of providing accurate classifications on unseen data from the same population. Removing branches from a tree generalises the model and reduces over fitting (Agresti, 2013).

We begin with an overly complex tree (Figure 16). The parameters have been set so that splits occur with three observations at each decision node (Table 14). To reduce the complexity of the tree, the tree can be pruned by increasing the 'complexity parameter'. By increasing this parameter we can see that the resulting models decrease in complexity, or fewer terminal nodes at the end of the tree (Table 14). Pruning is controlled in the rpart function by setting the minimum complexity parameter. Successive nodes from the root node have decreasing complexity parameters. The complexity parameter is a relative measure which comes from the amount of error improvement of a model by the addition of a node. The tree that was grown with no specified complexity parameter and three observations at each node has a complexity parameter of 0.01 at the terminal nodes (Table 14). As the complexity parameter is increased, the number of terminal nodes decreases, and the tree becomes less complex and more generalised.

Table 14: Toy data four modelled with a decision tree that has increasing complexity parameters to prune a tree with a minimum of three observations to split a node. The deeper model (the model without pruning and more terminal nodes) is a closer fit to the data and has a higher CV error than other models. Increasing the complexity parameter decreases the number of terminal nodes, and increases the AER and CV error, but the ratio of AER to CV error reduces.

Maximum terminal node complexity parameter	Number of terminal nodes	AER	CV Error
None specified (0.01)	11	0.75%	8.71%
0.03	8	3.01%	9.02%
0.06	5	9.02%	9.77%
0.12	3	17.29%	19.55%

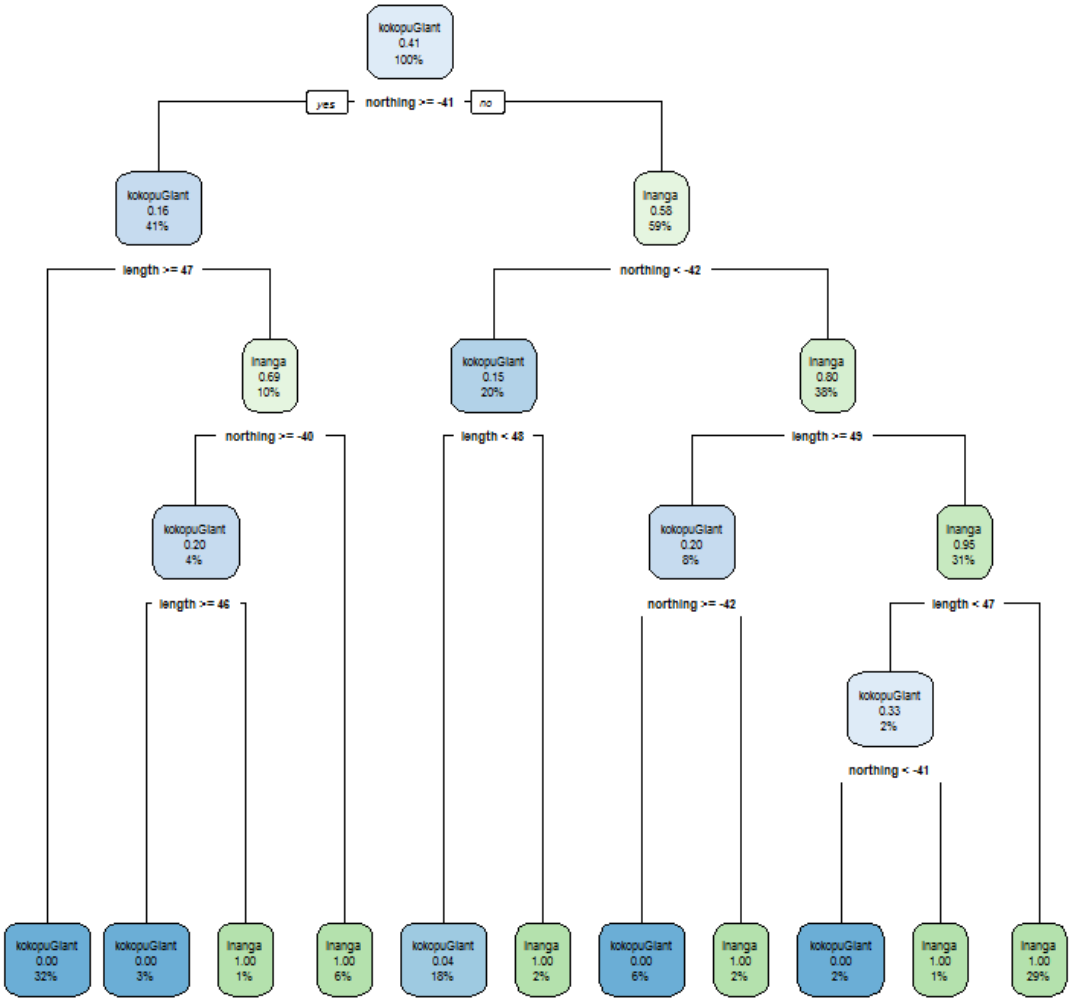


Figure 16: A decision tree for toy data four. This was created with a minimum of three observations at each node and no constraints on the complexity parameter. Eleven nodes result, and the terminal nodes have a complexity parameter of 0.01.

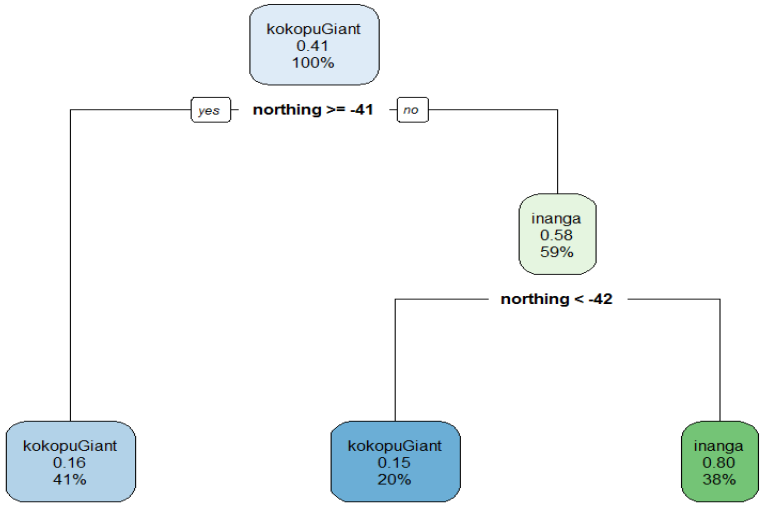


Figure 17: Decision tree with a minimum of three observations at each node, and minimum complexity parameter of 0.12. We now have three final nodes. This is a much more generalised tree than the one in Figure 16.

Prevalence of Species in Decision Tree

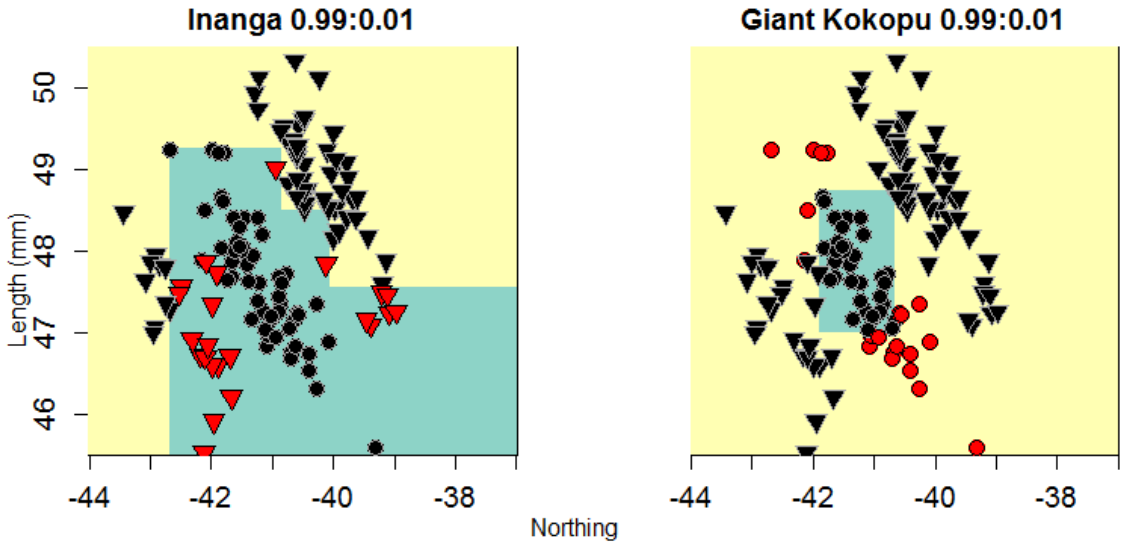


Figure 18: Decision trees from the toy data four. The models were created with altered prevalence for each species. The left panel shows a model created with the prevalence favouring īnanga. There are some giant kōkopu that have misclassified as īnanga, shown in red, but all īnanga have been correctly classified. The right plot shows the model performance with the prevalence weighted towards giant kōkopu. There are some īnanga that have been misclassified as giant kōkopu, shown in red. There are no misclassified giant kōkopu when the probability favours giant kōkopu.

Sometimes the species frequencies in the sample do not reflect population frequencies. Population frequencies may be considered by changing the species probabilities. Consider toy data four. If we

adjust the probabilities to favour īnanga 0.99:0.01 the fish with northing –41.9 and length 48.5mm the classification will be īnanga. Changing the prior probabilities to favour giant kōkopu 0.99:0.01 for the fish with northing -41.49, and length 48.07mm, the classification will still be īnanga (Table 15) (Figure 19).

Table 15: Confusion matrix for Decision Tree on all toy data four with altered priors. There are more observations misclassified as īnanga when the prior favours īnanga. More observations are classified as giant kōkopu when the prior is in favour of giant kōkopu.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV Error
toyData4 – Priors: Giant Kōkopu = 0.01, īnanga = 0.99			17.29%	36.84%
Predicted Giant Kōkopu	56	0		
Predicted īnanga	23	54		
toyData4 – Priors: Giant Kōkopu = 0.99, īnanga = 0.01			15.04%	15.79%
Predicted Giant Kōkopu	79	20		
Predicted īnanga	0	34		

Minimum Number of Observations that Create a Terminal Node

Changing the minimum number of observations required at each decision node alters the number of decisions that need to be made to classify an observation, and ultimately the number of ways a class can be selected. In the default rpart() tree, the minimum required number of observations to create a new decision node is 20 (Therneau & Atkinson, 2018). Consider toy data one. The default settings give the decision tree in Figure 13. There are five terminal nodes with two nodes for giant kōkopu and three for īnanga. If we decrease the required number of observations at each node to one, this means that one split will occur for every observation. The resulting tree has 28 terminal nodes with 14 ways to classify an observation as īnanga and 14 ways to classify an observation as giant kōkopu. Decision trees like this can end up quite long and complex even for small data sets (Table 16).

Table 16: Toy data four modelled with a decision tree that has increasing numbers of minimum observations at each node to grow the tree. The decision tree with fewer observations at every node was a closer fit to the data. For the decision tree with a minimum of three observations at each node the descriptive model AER is one order of magnitude lower than the CV error. As the minimum number of required observations at each node increases, the ratio between the descriptive AER and the CV error decreases substantially.

Minimum obs at each node	Terminal Nodes			AER	CV Error
	Total	īnanga	Giant Kōkopu		
3	28	14	14	0.75%	8.71%
10	9	4	6	3.01%	9.02%
25	5	2	3	9.02%	10.53%
50	4	2	2	16.54%	18.80%

To give example comparisons, consider toy data set four. If a tree model were to be grown with a minimum of three observations at each node, the resulting model would fit the data well (Figure 19) (Table 16) with one misclassified observation and a low AER, 0.75% (Table 16). CV error is considerably larger at 8.71% (Table 16). For the first model the descriptive model AER is one order of magnitude lower than CV error. As the minimum number of required observations at each node increases, the ratio between the descriptive AER and the CV error decreases considerably indicating less over-fitting.

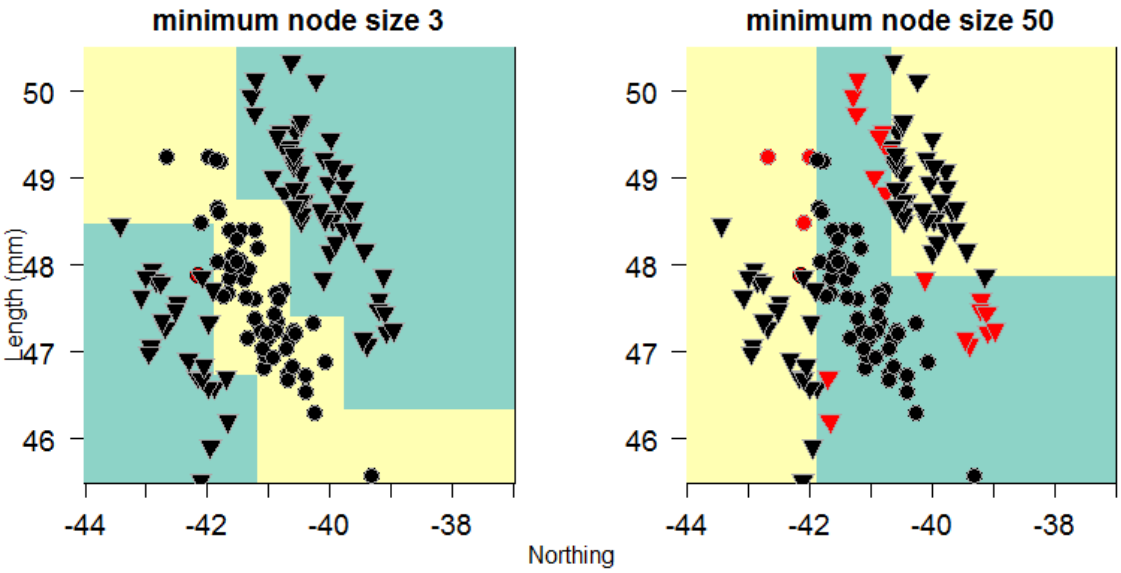


Figure 19: Decision tree plots for toy data four comparing a tree with only one observation at each decision node on the left with a tree that has at least 50 observations at each decision node on the right. The more observations required at each node fewer branches will populate the tree. The tree on the left has more decisions available than the tree on the right. There are some misclassified observations in the left panel, but more misclassified observations in the model with more observations at every node (right panel). Misclassified observations are indicated in red.

Alternatively, growing a tree with 50 observations at each node before a split can occur means that fewer splits will occur. Therefore, there will be fewer terminal nodes, so the decision tree will be shorter and less complex (Figure 19). A simpler decision tree means that it more likely to be able to be generalised to other samples from the same population. Decision trees with increasing numbers of observations required at each split have a decreasing number of terminal nodes (Table 16).

Maximum Number of Branches

Table 17: Toy data four modelled with a decision tree that has different maximum terminal nodes to grow the tree. The deeper tree, or the tree with more terminal branches is a closer fit to the data. The tree with 30 terminal nodes fits the data more closely, and classifies observations to the correct species more often with AER and CV error of less than 10%. The tree with 2 terminal nodes has a much higher AER and CV error.

Maximum terminal nodes	AER	CV Error
2	13.53%	23.31%
30	8.27%	9.02%

Increasing maximum number of branches increases the number of terminal nodes independent of the number of observations required for a split to occur. For example, take toy data four. If we increase the maximum number of branches to 50 then our tree does not change from the default tree. If we change the maximum number of splits to two, the tree has fewer terminal nodes (Figure 20).

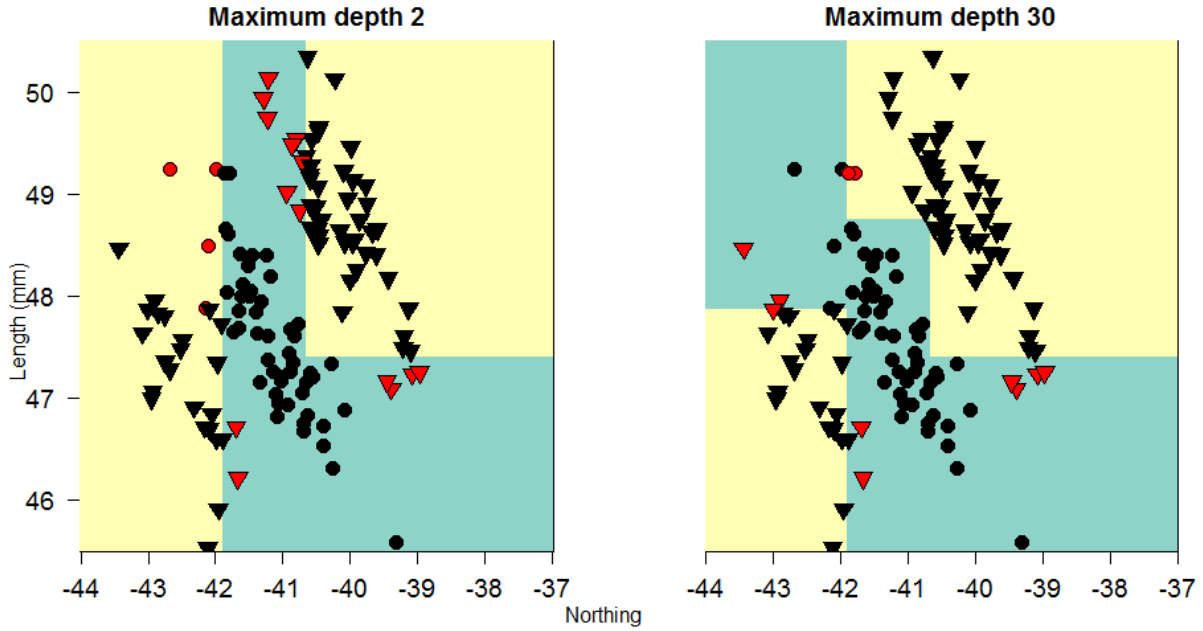


Figure 20: Toy data four modelled with a decision tree that has different maximum terminal nodes to grow the tree. Two terminal nodes is a less complex tree than the tree with 30 terminal nodes. A tree with has fewer terminal nodes and fewer decision branches (left panel) is less complex than the tree grown to have more branches (right panel). More observations are misclassified with the tree that has two terminal nodes. Misclassified observations are indicated in red.

Assumptions, Pros and Cons

Decision trees are non-parametric therefore there are no assumptions about the data.

Pros:

- Produces a set of questions that (often) does not require a computer to classify unseen observations (Figures 17 and 18) (James et al., 2013).
- Easy to interpret for non-statisticians.
- Able to take categorical predictors (Agresti, 2013).

Cons:

- Uses the greedy algorithm, which finds the local optimum for a variable, but not always the global optimum for all available variables (James et al., 2013).
- Can over-fit if appropriate restrictions are not placed on how the trees are grown (Agresti, 2013; James et al., 2013).

In general, the deeper the tree, the more complex the decision rules, and the closer the fit to labelled data but not necessarily to unseen data. Over-fitting the tree can mean that it is possible for the decision tree to have a leaf node for every observation. While this case may be useful for describing a dataset it is useless for predicting the classes of another dataset with the same variables (Friedman, Hastie, & Tibshirani, 2009).

Random Forest

Random forests grow an ensemble of trees and have the vote for the most popular class (Breiman, 2001). Data has the same properties as for decision trees.

To create the model:

1. Create some trees using the following algorithm:
    a. Draw a bootstrap sample from the data. A bootstrap sample is when a dataset is resampled with replacement to get a new sample of the same size as the original dataset
    b. Randomly select some of the variables
    c. Grow a tree using the algorithm from the decision tree method (Page 49) using only the randomly selected variables.
2. Output the ensemble of trees.

To make a prediction from this forest of trees, each tree in the forest takes a vote on which species the observation belongs to. The most popular vote is what the observation is classified to (Breiman, 2001).

Examples of Random Forest

Let us take an observation from toy data one with Northing -41.49, and Length 48.07mm. We would have to ask a computer to make a prediction using Random forest as the decision process is too complex to outline here. However the performance of random forest on different data sets is demonstrated in Table 18 and Figure 21. Each observation is correctly classified in all toy data sets. Even when the data are difficult to fit, Random forest has AER of 0%. However the CV error is different dependent on the complexity of the class mixing. Toy data one the mixing of the species is considerable, and CV error is 51.9%. Toy data two, the species are not mixed, but the there is still 2.5% CV error. Toy data three, the species are completely separate and the CV error is 0%. Toy data four, the species are separate but giant kōkopu are not in one group. CV error for toy data four is 4.51% (Table 18).

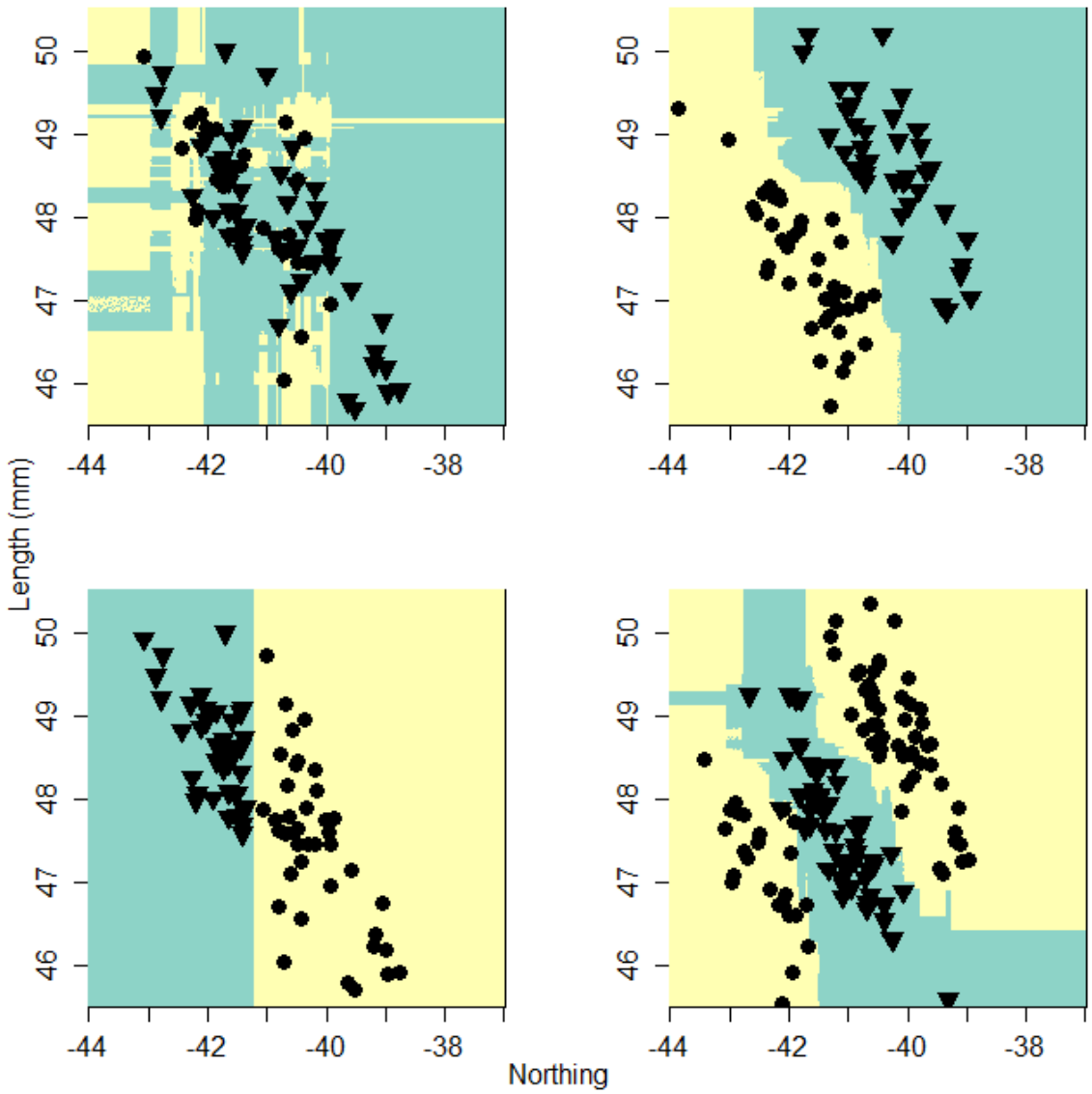


Figure 21: Random Forest performance on each toy data set. Each observation is correctly classified in all toy data sets. The top left and bottom right panels show how a model can over fit with data that would otherwise be difficult to model. Toy data one and toy data four have been described perfectly by these random forests.

Table 18: Confusion matrices for random forest classifier on all toy data sets. All random forests describe the data perfectly, but the CV error is different dependent on the complexity of the class mixing. Toy data one the mixing of the species is considerable, and CV error is high. Toy data two, the species are not mixed, but the there is still 2.5% CV error. Toy data three, the species are completely separate and the CV error is 0%. Toy data four, the species are separate but each species is not in one group. CV error for toy data four is 4.51%.

	Simulated Giant Kōkopu	Simulated Īnanga	AER	CV error
toyData1			0.00%	51.90%
Predicted Giant Kōkopu	25	0		
Predicted īnanga	0	54		
toyData2			0.00%	2.50%
Predicted Giant Kōkopu	40	0		
Predicted īnanga	0	40		
toyData3			0.00%	0.00%
Predicted Giant Kōkopu	41	0		
Predicted īnanga	0	38		
toyData4			0.00%	4.51%
Predicted Giant Kōkopu	79	0		
Predicted īnanga	0	54		

Altering Random Forest Controls

If we compare a random forest to a decision tree most of the same controls that are available to decision trees can be manipulated with random forests for each random tree. All controls are going to be the same for each random tree grown in the random forest.

Species Prevalence in Random Forest

Sometimes the class frequencies in the sample do not reflect population prevalence. Population frequencies may be considered by changing the species prevalence. Consider toy data four. Altering the prevalence made no difference the AER as it was 100 % (Table 19). CV error showed that the model does not have perfect prediction, and is over fitting. The CV error was better for the model that favoured giant kōkopu (Table 19).

Table 19: Comparing the performance of random forests with decision trees that were generated using toy data four with altered species prevalence. There are more observations misclassified as īnanga when the prevalence favours īnanga. The CV error for random forest is less for random forest than for decision tree. More observations are classified as giant kōkopu when the prevalence is in favour of giant kōkopu. Both models are over fit and have AER of 0%. The CV error is smaller when the prevalence of īnanga is 0.01 for decision tree and for random forest, and much smaller for random forest.

Prevalence of īnanga (as a proportion)	Decision Tree		Random Forest	
	AER	CV error	AER	CV error
0.99	17.29%	36.84%	0.00%	7.52%
0.01	15.04%	15.79%	0.00%	3.01%

Minimum Number of Observations that Create a Terminal Node

Requiring more observations at each node means each tree fits the observed data less well, and reduces model accuracy. Let us use toy data set four for an example. As the number of observations at each decision node increases the AER increases, as does the CV error. Random forest model is an improvement over the decision tree with the same number of observations at each node, except when there are 50 observations required at each decision node (Table 16). Decreasing the number of nodes required at each decision node in a random forest tree makes a more complex forest of trees.

Table 20: Toy data four modelled with a random forest that has different numbers of minimum observations at each node to grow each tree. As the number of observations at each decision node increases the AER increases, as does the CV error. Random forest model is an improvement over the decision tree with the same number of observations at each node, except when there are 50 observations required at each decision node

Minimum observations at each node per tree	Decision Tree		Random Forest	
	AER	CV error	AER	CV error
3	0.75%	8.71%	0.00%	4.51%
10	3.01%	9.02%	0.75%	5.26%
25	9.02%	10.53%	5.26%	9.02%
50	16.54%	18.80%	21.80%	28.57%

Maximum Number of Terminal Nodes

More terminal nodes mean a closer fit to the data for each tree. Increasing the number of terminal nodes available increases the complexity of each tree in the random forest, so the model is closer fit to the data. Similar to decreasing the required number of observations at each decision node, increasing the number of terminal nodes each tree in a random forest can be grown to will make for a more complex forest of trees. For example, take toy data four. If we compare the random forest model performance with altered maximum nodes it can be seen that the more complex forest with a maximum of 30 terminal nodes performs better than the less complex forest with maximum two terminal nodes (Table 21). Comparing random forest model performance to decision trees with the

same complexity, we can see that the more complex forest outperforms the more complex tree, but the less complex tree outperforms the less complex forest (Table 21).

Table 21: Toy data four modelled with a random forest that has different numbers of maximum number of terminal nodes. As the number of maximum terminal nodes increases the AER increases, as does the CV error. Random forest model is an improvement over the decision tree with the same number of maximum terminal nodes.

Maximum terminal nodes per tree	Decision Tree		Random Forest	
	AER	CV error	AER	CV error
2	13.53%	23.31%	27.82%	32.33%
30	8.27%	9.02%	0.00%	3.00%

Number of Trees Generated in a Forest

The number of randomly generated trees can be changed. Take for example, toy data four. If we were to increase the number of trees in a random forest and do 10 fold cross validation on each forest we can assess the model performance as the number of trees changes. Increasing the number of random trees increase the overall model accuracy, reducing the number of random trees reduces the model accuracy, and increases the CV error (Figure 22). A random forest with one tree does not do better than a decision tree most of the time (Figure 22). The addition of more trees increases the sensitivity to a point. In this example, ~5% CV error is consistently reached after 30 trees are generated (Figure 22).

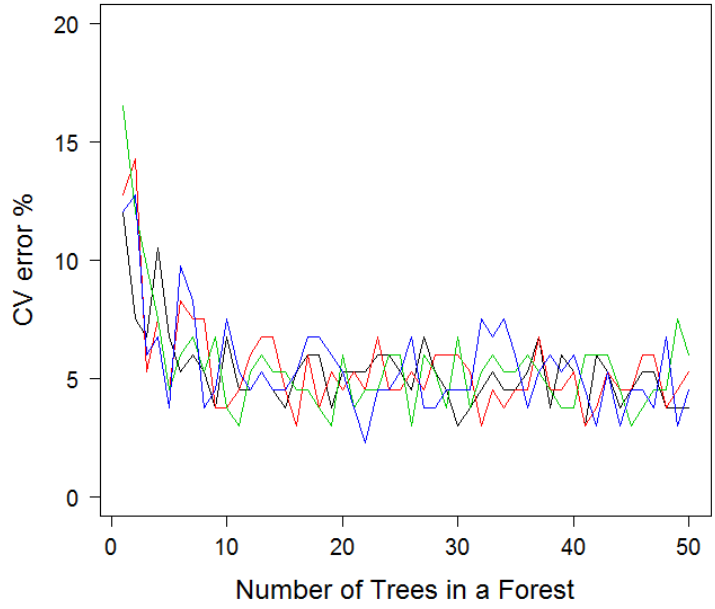


Figure 22: CV error of four random forest traces with increasing numbers of trees for toy data four. Four simulations of random forests were run with an increasing number of trees included in each random forest. Increasing the numbers of trees grown in each forest decreases the CV error of the model. Once more than approximately 30 trees have been generated the CV error is consistently ~5%.

Pruning Trees

Pruning of trees within a random forest does not occur because it is not necessary (Breiman, 2001). Each tree is generated from an uncorrelated out of bag sample. The algorithm produces many trees that are uncorrelated 'out of bag' samples. Each sample is a bootstrap aggregation of sub-samples (bagged sample) from labelled data. Random forest removes the correlation of bagged trees by choosing the variables to split on before each tree is grown. The number of variables chosen does not appear to affect how good classification is. However, the best results seem to come from selecting one or two variables. The default value for the number of variables selected is $\lfloor\sqrt{p}\rfloor$ where p is the number of available variables. As random forests are random, each time a forest is grown it will fit the data slightly differently and yield a slightly different error rate.

Assumptions, Pros and Cons

Like decision trees, random forests do not need any distribution for the algorithm to model the data as each tree is non-parametric.

Pros

- Very high accuracy
- Can take categorical variables

Cons

- Unable to be used for classification without a computer.
- Likey to overfit.

Diversity Measures

The Shannon Diversity is a measure of entropy or evenness of species abundances. It is calculated using Equation 17.

$$S = -\sum_{i=1}^{n} p_i \ln(p_i) \quad [17]$$

Where p is $\frac{n_i}{N}$ and $n_i$is the number of observations in class i and N is the number of observations in total. The lower the Shannon entropy the less diverse the data is in terms of species. Shannon diversity has no upper bound, but it is rare for it to be above 4.0.

The Simpson index is a measure of the diversity that indicates if one species if one species is dominant. This index is calculated using Equation 18.

$$D = 1 - \frac{\sum_{i=1}^{n} n_i(n_i-1)}{N(N-1)} \quad [18]$$

If one species is dominant the Shannon Diversity will less than 0.5, and the Simpson Index will be close to 1. If all species are present in near equal abundances, the Simpson index will be lesss than 0.5, whereas the Shannon diversity will be large, dependent upon the number of species in the system. A Simpson's index can only be 1.0 when there is one species present. If there are equal abundances of species, where the number of species approaches infinity, the Simpsons index would be close to zero.

Table 22: Diversity measures for all toy data sets. These all have approximately similar diversity and dominance measures. Toy data two has the same prevalence of īnanga and giant kōkopu so the Shannon diversity is 1.0.

Data set	Shannon Diversity	Simpson Index
Toy data one	0.901	0.562
Toy data two	1.000	0.494
Toy data three	0.999	0.494
Toy data four	0.974	0.514

Methods summary

In summary, MLR works well when species distributions are linearly separable, and not so well when the species distributions are mixed, or when one of the species has two groups like toy data four (Table 23). LDA and QDA have similar properties to MLR, but QDA did a better job at picking out the two groups of giant kōkopu from toy data four than LDA and MLR. Naive bayes did not classify the species well in toy data one, but performed well when the species distributions were linearly separable, as in toy data two and toy data three. The distibution of īnanga in toy data four was described realtively well using Naive Bayes, but the AER and CV error were still close to 20%. Decision tree was unable to pick out the species in toy data one with AER of 26.58%, and CV error of nearly 50%. Decision tree predicted classes well for toy data three with AER and CV error 0%. Toy data two were not classified with 100% accuracy, even though there was a clear separation between the two species. Toy data four were classified with AER and CV error of less than 10% using decision tree. Random forest over fit every toy data set with AER of 0%. CV error for random forest model gave better indication of how good random forest was at classificatrion. Toy data one species were very mixed, and the CV error was 51.90%. Toy data two species were linearly separable, and random forest CV error was 2.5%. Toy data three species are separable by Northing. CV error for random forest with toy data three was 0%. Toy data four, giant kōkopu are in two separate groups, but very separate from īnanga. Random forest CV error was 4.51% (Table 23).

Table 23: Toy data set properties with apparent error rates (AER) for model and the CV error for each model. All figures have been given to four decimal places. No classification method classified toy data one well in cross validation. Toy data two was well classified across all methods, except decision tree where CV error was 13.75%. Toy data three was classified well with all methods. Toy data four had mixed results. MLR, LDA and Naïve Bayes all classified toy data four poorly. Decision tree classified toy data four with higher accuracy than MLR, LDA and Naïve Bayes. QDA classified species in toy data four with CV error of 6.01%. Random forest was over fit to every toy data set as evidenced by the 0% AER and higher CV error. In particular, toy data set one had CV error or 51.9%

		MLR		LDA		QDA		Naïve Bayes		Decision Tree		Random Forest	
Set	Unique property	AER	CV Error	AER	CV Error	AER	CV Error	AER	CV Error	AER	CV Error	AER	CV Error
1	Species are indistinct from each other	31.64%	34.17%	31.64%	34.18%	32.91%	35.44%	31.65%	36.71%	26.58%	45.57%	0.00%	51.90%
2	Species separable by a linear combination of northing and length.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.25%	13.75%	0.00%	2.50%
3	Species separable by northing	0.00%	0.00%	1.27%	2.53%	1.27%	2.53%	1.27%	1.27%	0.00%	0.00%	0.00%	0.00%
4	Species clearly separated, but one species has two distinct groups	42.11%	42.11%	42.11%	42.86%	3.76%	6.01%	17.29%	18.80%	8.27%	9.02%	0.00%	4.51%

Section 3: Data

Data descriptions

This project used morphological and environmental data that was collected as part of Mark Yungnickel's master's project (Yungnickel, 2017). All variables are described in Table 24. Data were collated from sites across New Zealand (Figure 23) between July and December 2015 by a mixture of recreational fishers that donated a portion of their catches, and data collection data for this project.

Table 24: Variable descriptions. 'depth' had the highest proportion of missing values, followed by 'lengthFro', then 'weigthFro'. No other variables had missing values.

Variable	Description	Variable Type (r variable type, data type)	Range/Possible values	Units	Proportion missing (3dp)
species	Fish species	as.factor nominal	kokopuSJ', 'kokopuBand', 'kokopuGiant', 'koaro', 'inanga'	-	0
fishID	Observation ID	as.factor nominal	1 to 17545	-	0
date	Date sample was taken	as.date	July to December 2015	-	0
region	Spatial region	as.factor nominal	Appendix 1	-	0
riverID	Name of watercourse sampled.	as.factor nominal	Appendix 1	-	0
lengthFro	Frozen length from nose to tail tip	as.numeric continuous	35.44 , 59.77	mm	0.033
weightFro	Frozen wet weight of the whole animal	as.numeric continuous	0.12 , 1.08	grams	0.012
depth	Distance between insertion of dorsal fin to insertion of anal fin (Figure 4)	as.numeric continuous	2.00 , 7.40	mm	0.096
lat	Watercourse mouth Latitude	as.numeric continuous	-46.60 , -36.42	°	0
lon	Watercourse mouth Longitude	as.numeric continuous	167.61 , 178.01	°	0

The country was divided into regions based loosely on district council boundaries. As district council boundaries tend to be defined by rivers, each river was assigned to only one region (Figure 23). Each

region had between 1 and 14 rivers (Table A1). The decision to define regions was made to characterise the spatial heterogeneity of species composition and within species diversity of spatially close populations. Latitude and longitude were estimated from district council records. In-stream conditions were recorded, but not used for the purposes of this project.

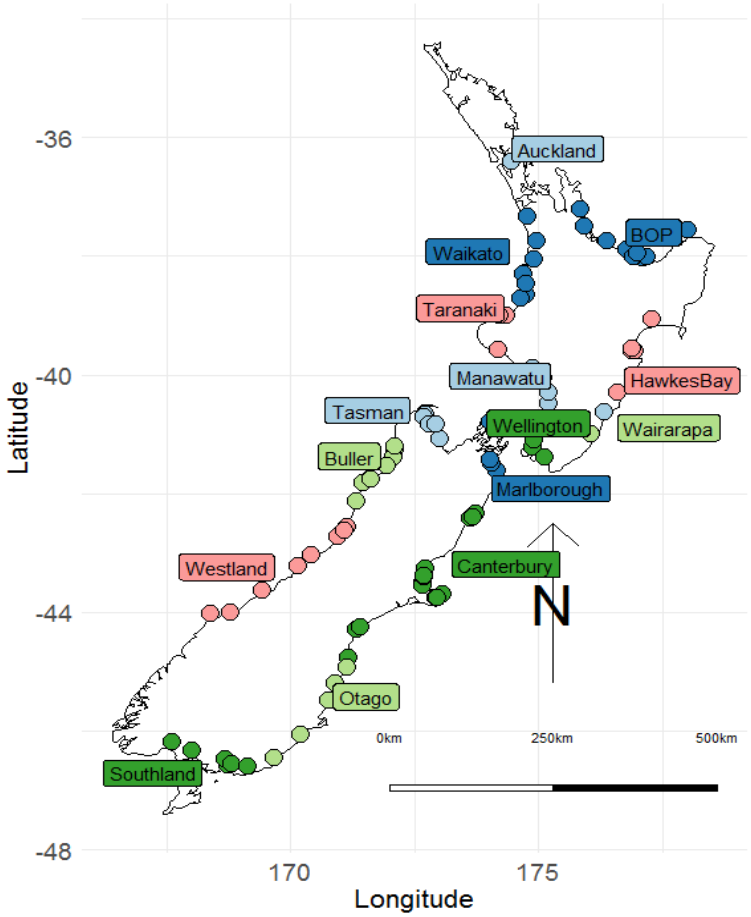


Figure 23: Map of region locations. Each dot on the map is an approximate location of a river mouth where samples were collected from. The colour of each dot indicates the region it was assigned to.

Measures of morphology were wet length, from nose tip to tail tip in millimetres, wet body weight in grams, and body depth in millimetres. Depth is a new measure (Figure 24). The depth measure is being tested to see if it can capture the diversity of the dorsal fin insertion point as it relates to the anal fin insertion point. Depth was calculated from microscope measurements. The error in depth measurement is ±0.01mm. Length was measured using electronic callipers (± 0.01 mm). Whitebait were weighed on an electronic balance (± 0.001 grams). Non-morphological measures were the date of capture, and an estimation of the latitude and longitude at the mouth of the river the fish was captured. Latitude and longitude were estimated using regional council data. Measurement errors in the latitude and longitude for each location had meant that some of the sampling sites appeared to be in the ocean (Figure 25). However, for the purpose of this study the sampling locations did not

need to be precise. Other non-morphological measures included environmental conditions but these were included in any analyses.

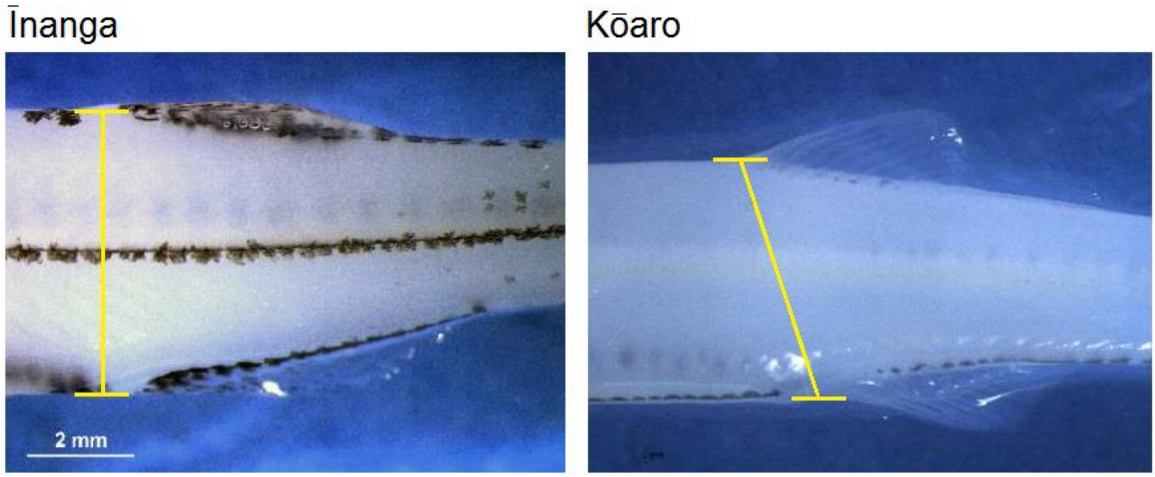


Figure 24: Showing the difference in the depth between fish species. Species with the same body length and weight will have a different depth when the dorsal fin insertion point is in line with the anal fin insertion point, compared to offset in front of the anal fin insertion point. The dorsal fin is at the top of fish, and the anal fish is at the bottom of the fish. On īnanga the insertion point of the pectoral fin tends to be in line with the anal fin forming a perpendicular line between the line of the insertion points and the lateral line of the fish. On the kōaro, the insertion point of the pectoral fin tends to be offset from the anal fin.

There were 17548 observations recorded with ~3% of those missing some measurements (Table A2). Missing measurements were associated with region, and with sampler (Table A2). There were no missing dates or spatial measures (Table 24). Depth was the most commonly missed measurement. For the whole country, there were ~10% of depth measures missing, ~3% of frozen total length measures missing, and ~1% of weight measures missing (Table 24). The proportion of missing data was not consistent for all regions or species (Appendix 2). Observations from Auckland had ~79% of depths from banded kōkopu missing missing and Canterbury had 10% depths missing from īnanga. Pre-frozen measurements of fish were taken, but were discarded for this project as there were few fresh measures. Fresh and frozen measures have a high linear correlation (Yungnickel, 2017) so frozen measures can be used in place of fresh measures for the purposes of this project.

Short jaw kōkopu was targeted in some rivers, but only two fish were identified as shortjaw kōkopu. There was no attempt to target species in any other rivers. As such, the sample prevalence of each species is taken as representative of the population prevalence of each species in the fishery. The fishing methods used by whitebaiters make bycatch common (McDowall, 1965). Bycatch is where other fish are caught with the target species. Some bycatch species were included in samples. These are not galaxiids and were discarded from the final analyses. I was not able to find any studies on

whitebait catchability. Not all species were detected in every region. The number of observations at each river was not the same (Figure 24).

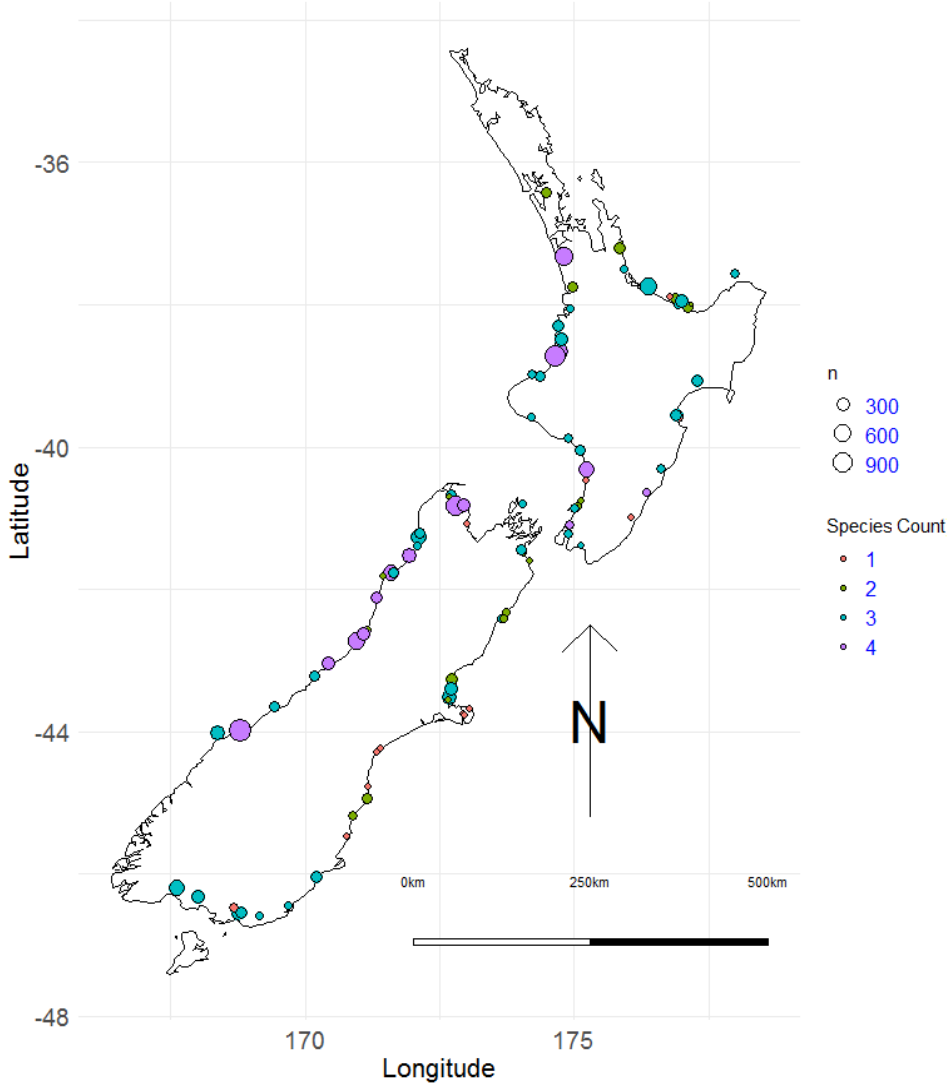


Figure 25: Map with the number of observations from each river, and the number of species found in each river. Maps created in R using package ggmap (Kahle & Wickham, 2013). There were more species observed on the West Coast then on the East Coast. More fish were taken from West Coast rivers than East Coast rivers.

Sampling effort was not specifically recorded. One way of assessing sampling effort would be the number of sampling occasions at each watercourse. From data we can infer that sampling effort was not consistent across all rivers or all dates (Table A1). Each sampling event by each whitebaiter had at least 40 observations (Table A1). The average number of dates that a river was sampled was three times, with a maximum of 16. The average number of observations taken from each river was 194.4 with a maximum of 1190 and minimum of two.

Exploratory data analysis

Data covered whitebait sampling from July to December (Figure 26). Each species was caught in different quantities across the sampling period (Figure 26). Most fish were caught during November. November is in the middle of whitebaiting season. Īnanga was the most common species throughout the season. There were two observations positively identified as short jaw kōkopu. As such, it would be ill-advised to draw inferences from these, so they have been removed from the data.

The distribution of the weights, depths, and lengths was different for each species and average body size is different between regions (Figures 27, 28, 29). Banded kōkopu have the shortest mean length in all regions they are present. Īnanga had the highest mean length, except in Waikato, Hawkes Bay, Tasman, Buller, and Otago. Where they are present, banded kōkopu have the smallest mean depth. Kōaro have the largest mean depth in almost all regions. In Westland the largest mean depth is giant kōkopu. Īnanga has the largest mean depth measurement in Auckland only. Banded kōkopu have the smallest mean weight where they are present. Kōaro have the heaviest mean weight in all regions where they are present.

We calculated Shannon Diversity and Simpson's Index for each region (Keylock, 2005) (Table 25). A larger means more species in the community. Systems with more than two species that had a Shannon diversity smaller than 0.5 tells us that there is one dominant species. Shannon diversity of over 1.0 in a system with more than two species in the population indicates that species abundances are even.

Simpson's Index is a diversity measure that indicates if there are some species more dominant than others. Simpsons indices close to 1.0 indicates that when more than one species is present that they are not very abundant. It is important to compare both Shannon diversity and Simpson's index between regions as they quickly indicate regions with high diversity, and regions with one very dominant species.

Shannon Diversity and Simpson's index were calculated the whole nation, and for each region. Shannon diversity is a measure of entropy or evenness of species. The Simpson index is measure of dominance. The national Shannon Diversity was 0.822. The only region that had Shannon Diversity over 1 with fewer than four species was Taranaki. The Simpson's Index for Taranaki was 0.45. Wairarapa has a Shannon index of 0, and a Simpson's index of 1. Tasman had the lowest Simpson's Index (0.39) and four species detected. Canterbury had the lowest Shannon Diversity, and the

highest Simpson Index. Buller was the most diverse region with Shannon diversity of 1.636, and Simpson's index of 0.34.

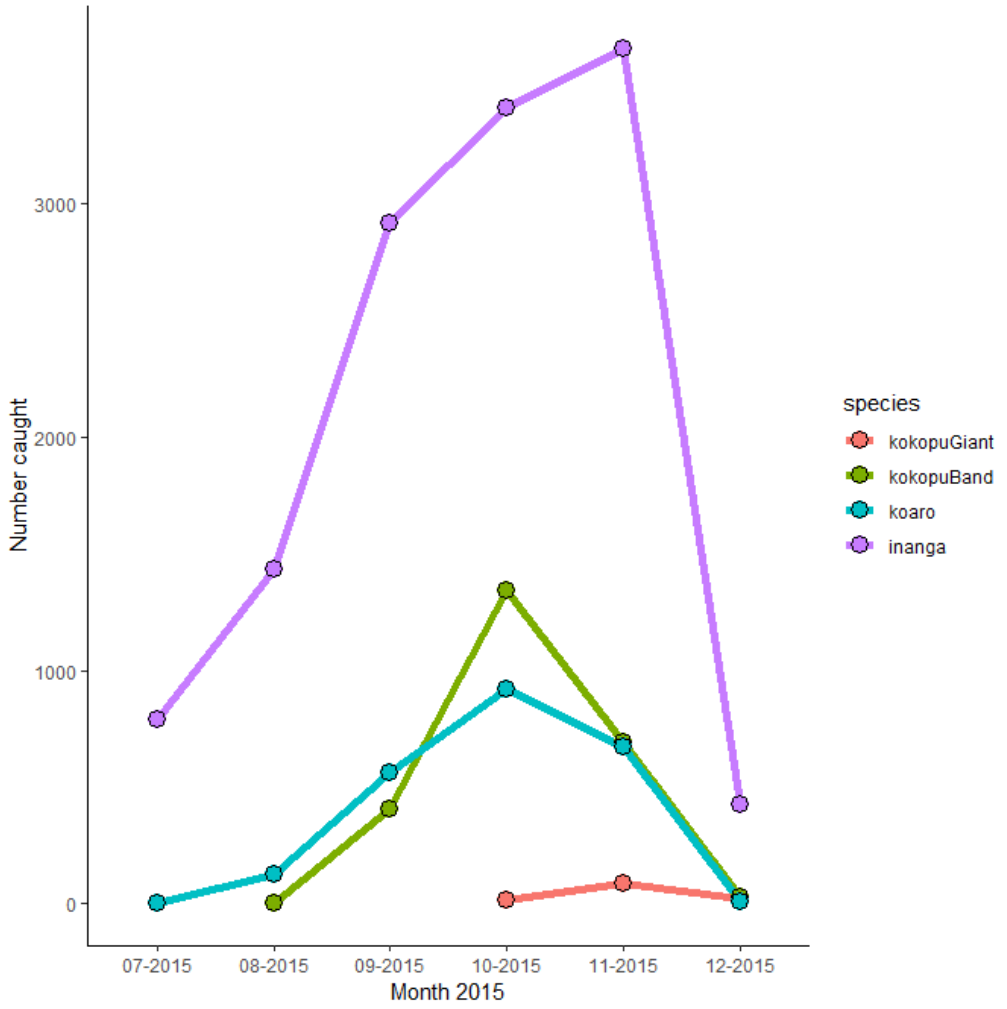


Figure 26: Number of each species that were sampled each month. The largest volume of fish was sampled in October. The most īnanga was caught in November. There were no giant kōkopu detected in July, August, nor September. The most giant kōkopu were detected in November. Banded kōkopu were not detected until August. Fewer fish were sampled during December than any other month.

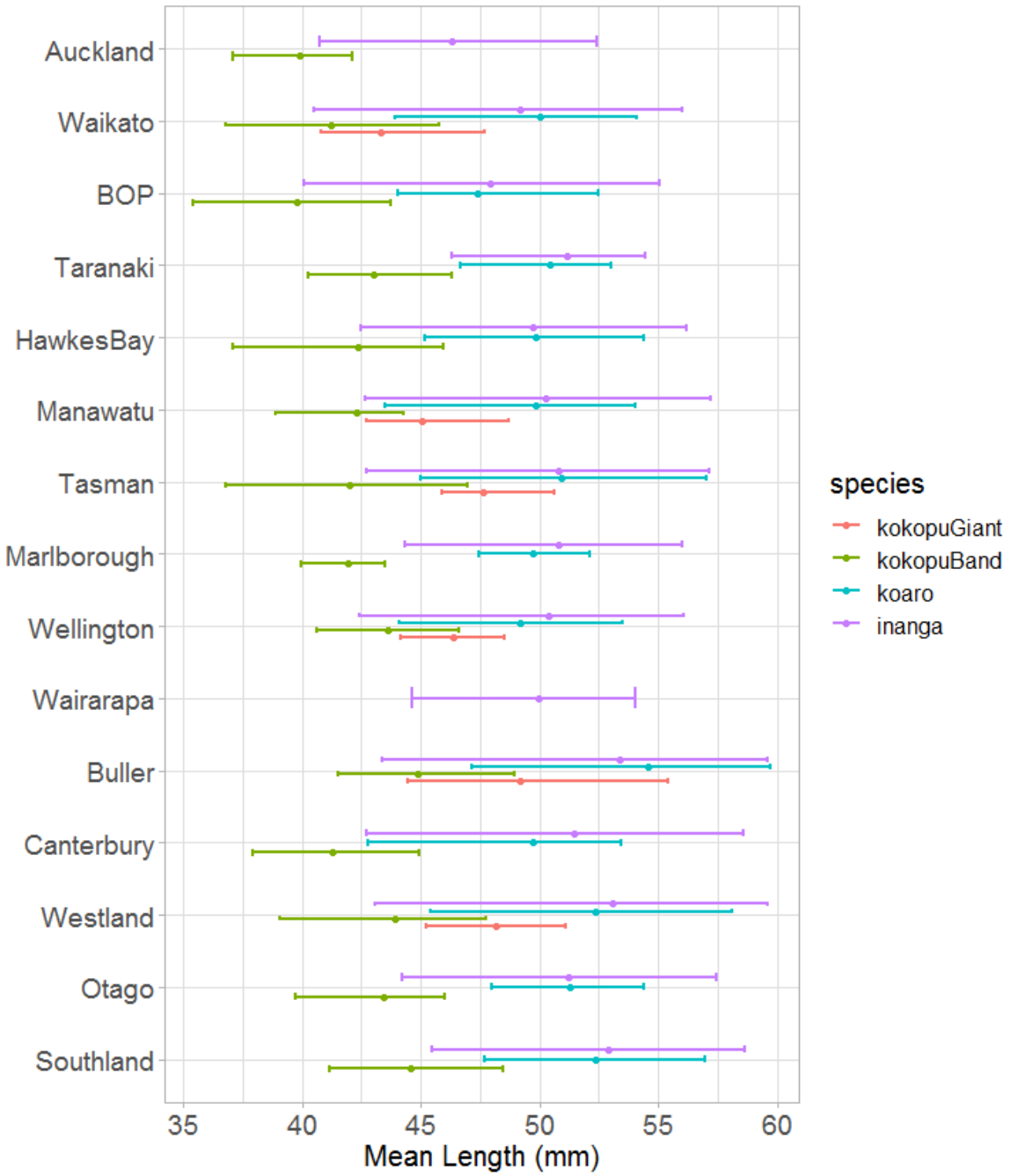


Figure 27: Mean length of all species for each region with range. Where they are present, banded kōkopu are the shortest fish. Giant kōkopu are near second shortest in the regions where they are present. Īnanga and kōaro have similar lengths.

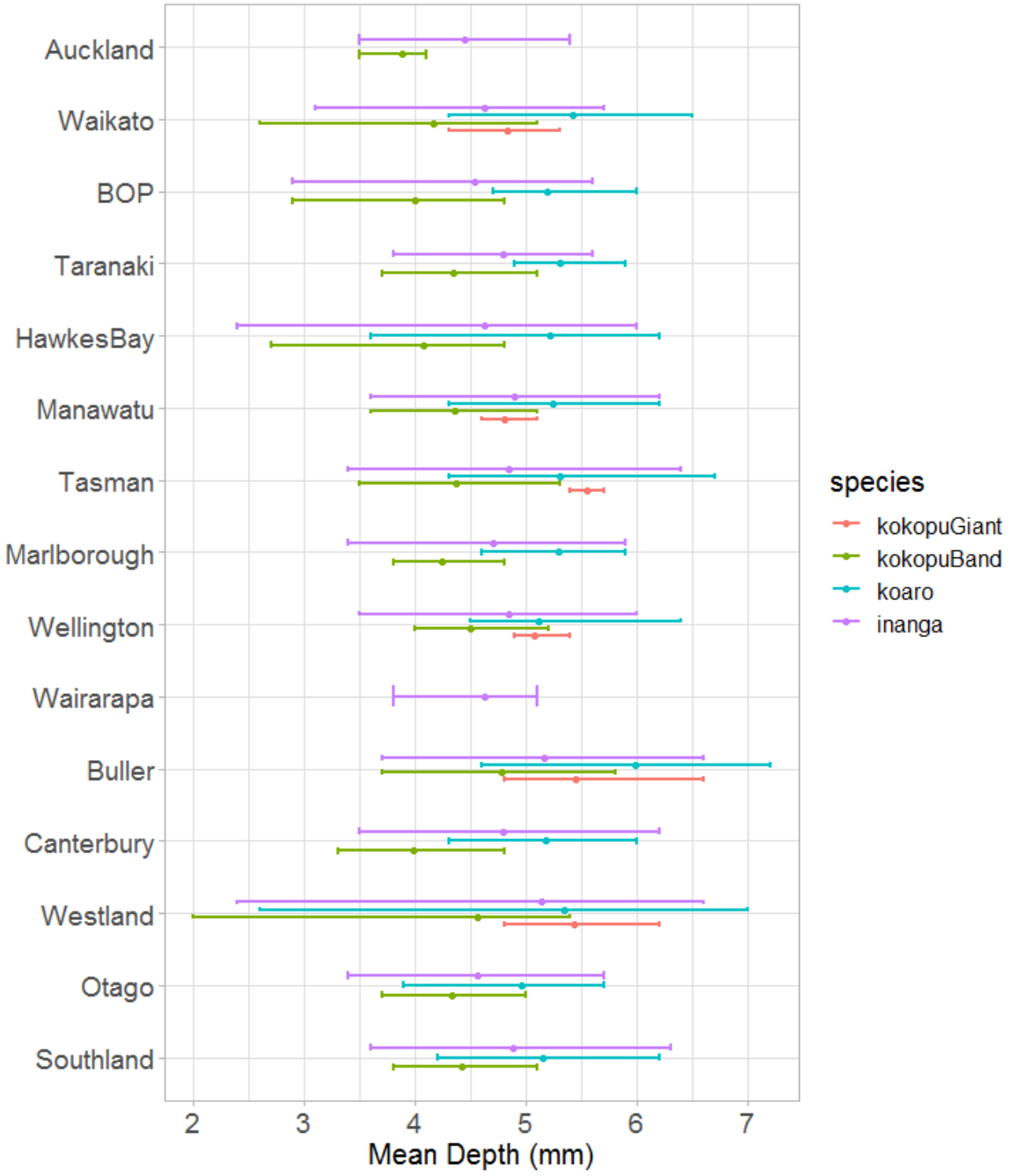


Figure 28: Mean depth for all species for each region with range. The depths are similar between species, within regions. Westland had the widest range of depths for all species. The only species detected in Waiararapa was īnanga.

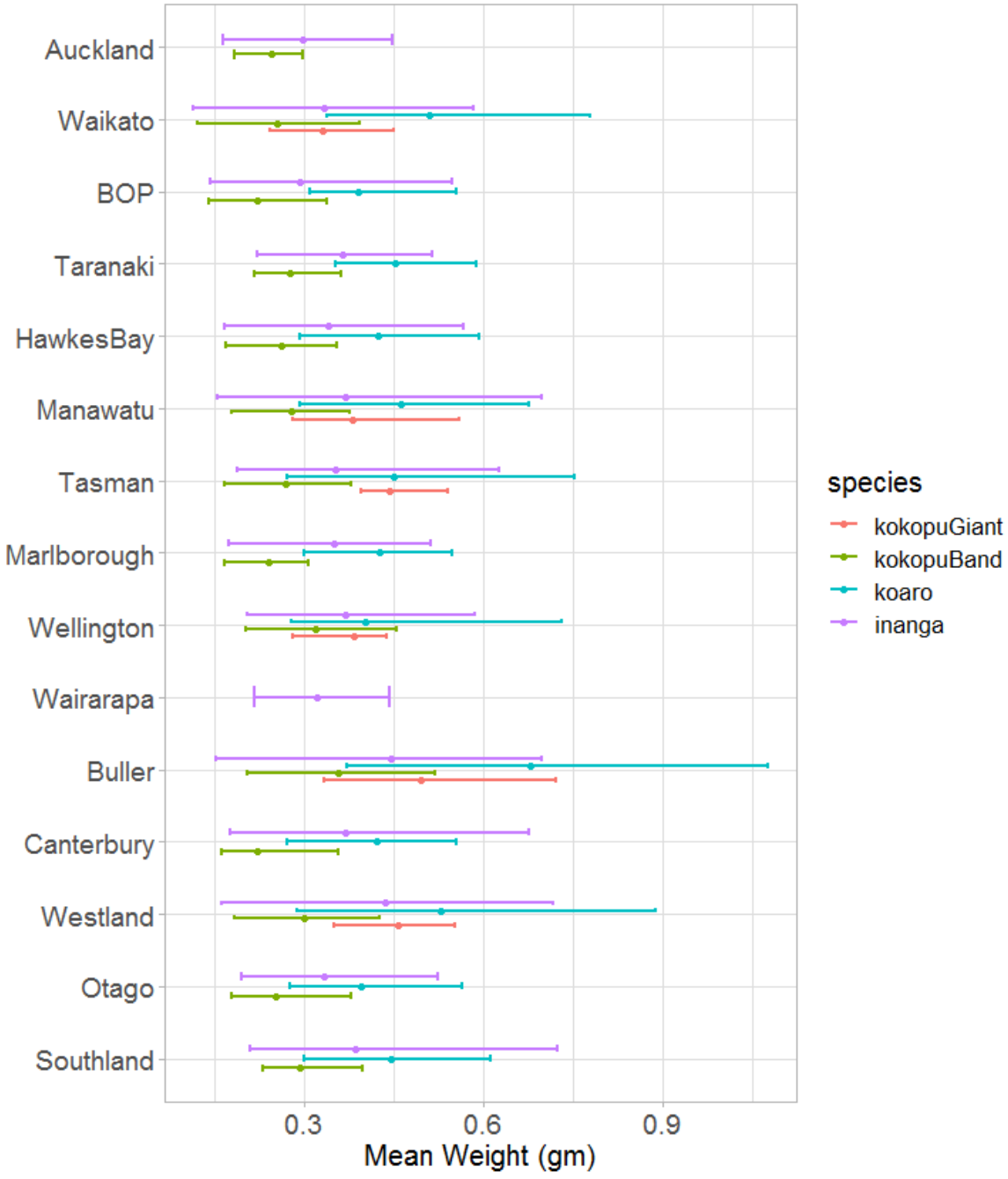


Figure 29: Mean wet weights for all species in all regions with range. Banded kōkopu are the lightest fish in regions they are present. Kōaro are the heaviest fish in Tasman, Wellington, Buller, and Westland. Buller had the widest range of weights.

Table 25: Shannon diversity and Simpson's index for National data, and for each region. In Wairarapa there was only īnanga detected. Shannon diversity for Wairarapa was 0.00 and Simpson's Index was 1. Buller was the most diverse region with Shannon diversity of 1.636, and Simpson's index of 0.34.

Region	Shannon Diversity	Simpson's Index
National	0.822	0.56
Auckland	0.822	0.65
Waikato	0.904	0.65
BOP	0.704	0.75
Taranaki	1.273	0.45
Hawkes Bay	0.642	0.78
Manawatu	1.213	0.54
Tasman	1.488	0.39
Marlborough	0.726	0.75
Wellington	1.357	0.48
Wairarapa	0.000	1.00
Buller	1.636	0.34
Canterbury	0.447	0.86
Westland	1.335	0.49
Otago	0.835	0.696
Southland	0.787	0.71

There is greater with region variance than between for all morphometric measures, especially length (Table 26). The date of capture for all species is similar (Table 26). The earliest mean migration time was īnanga. The latest mean migration time was giant kōkopu. Īnanga and kōaro had similar mean length.

Table 26: Between region, within region, and total variance for weight, length and depth. There is more variation within regions than between regions for all morphometric measures.

Variable	Total variance	Within region variance	Between region variance
Length	16.7509	178.00	4.060
Weight	0.3784	0.10	0.007
Depth	0.3416	3.38	0.152

When viewed as joint density distributions, the densities of morphometric measures show that there are differences between distributions of each species. Īnanga and kōaro had similar joint density of length and depth. Weight and depth joint densities appear to be similar for all species (Figure 29).

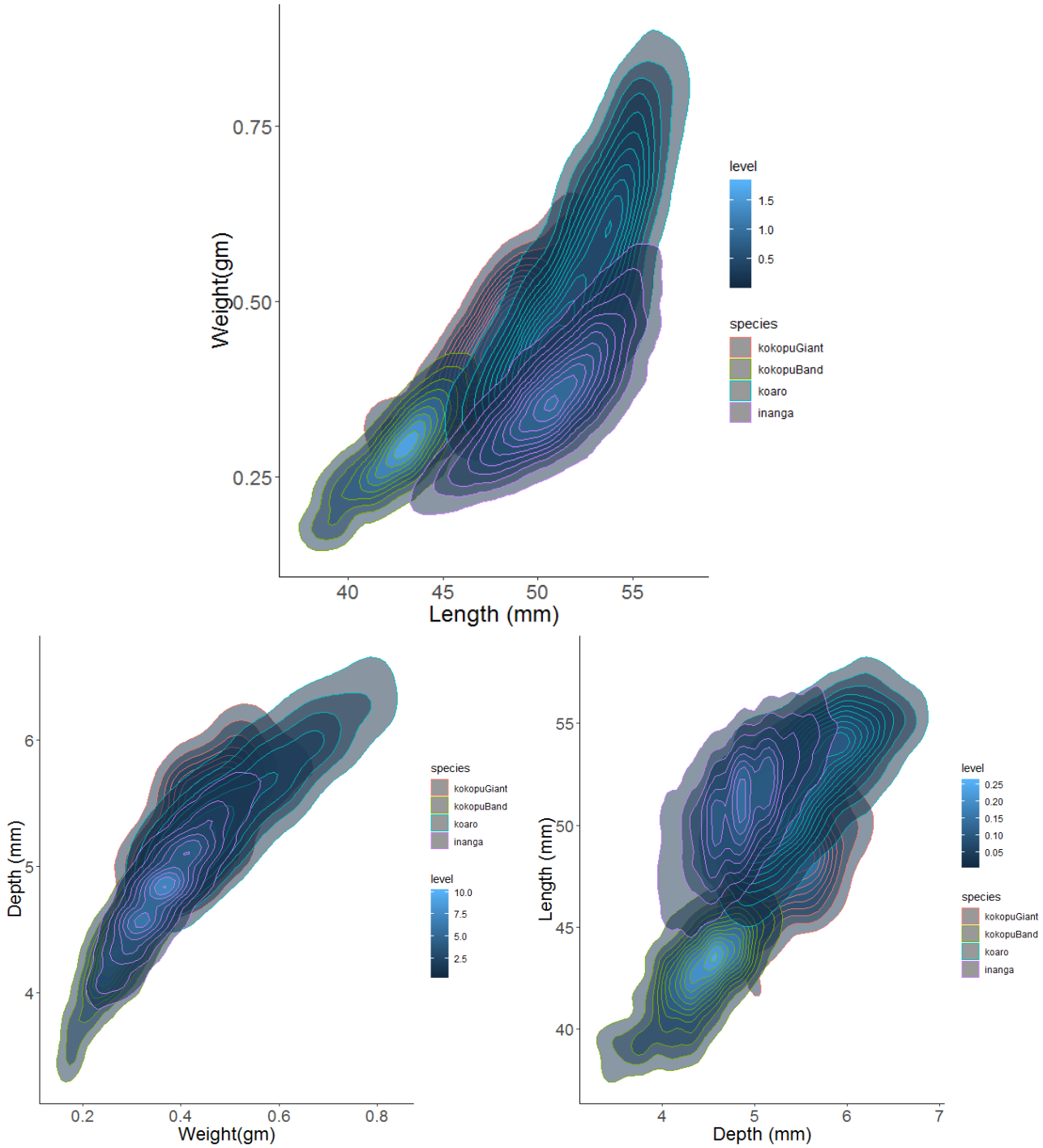


Figure 30: Densities of morphometric measures by species. Densities for weight v length appears to be different for each species. Densities for weight v depth appear to be similar for all species. Densities for length v depth appear to be similar for īnanga and kōaro, but different between giant kōkopu and banded kōkopu.

Morphometrics appeared to change for each species over the season (Figure 30). The month with the largest mean length fish is August, except for giant kōkopu (Panel A). The highest mean length for giant kōkopu is November. The month with heaviest mean fish for īnanga and kōaro is September. The month with heaviest mean banded kōkopu is October, and giant kōkopu is

November. The month with greatest mean depth īnanga and kōaro is September. The greatest mean depth banded kōkopu was in October, and for giant kōkopu was December.

The mean sample date of each species reflects the pattern from Figure 26. The approximate dates that the species were caught in the highest abundances were close to the mean capture dates of each species (Table 27). Īnanga were caught in highest abundance in November which is much later than the mean date of 5 October.

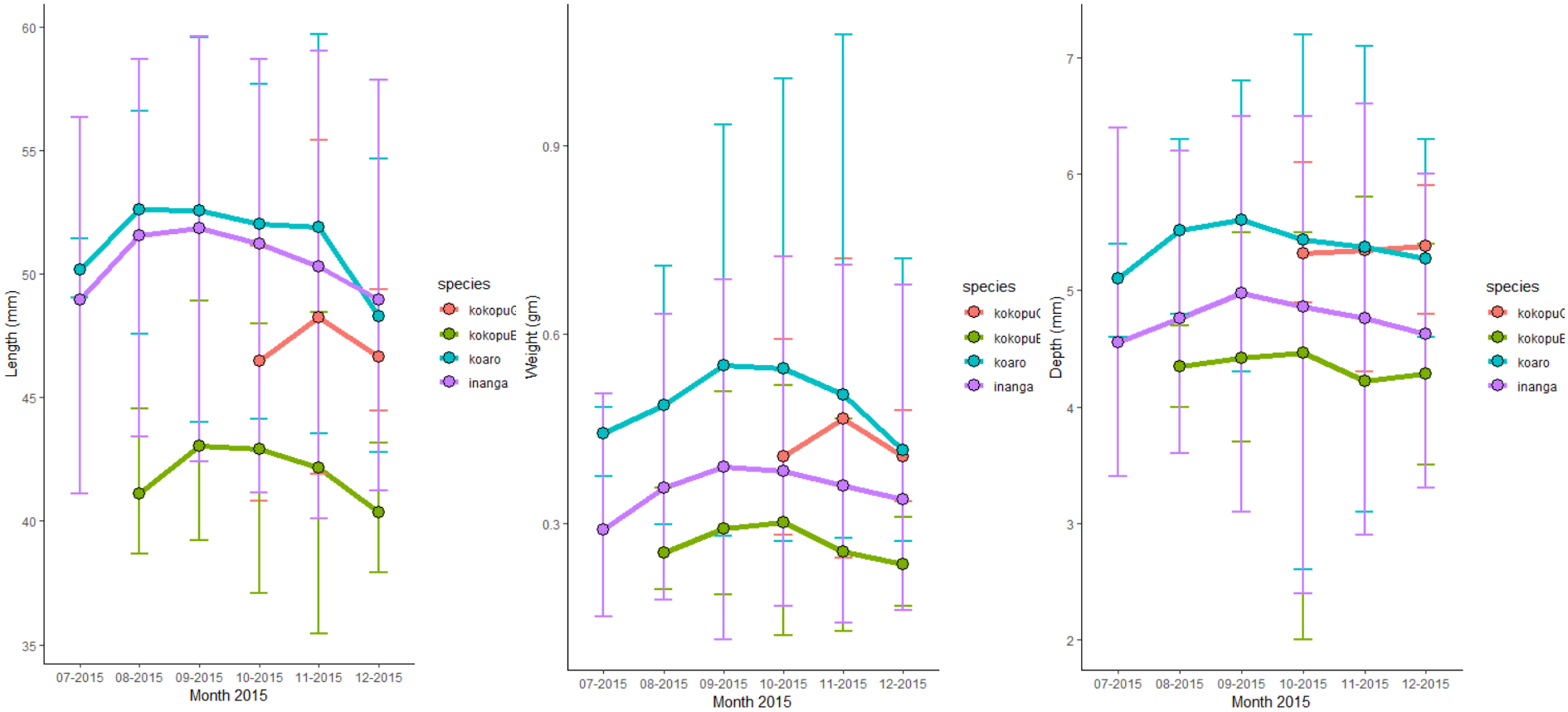


Figure 31: Mean morphometrics of each species by month with bars depicting range. The month with the largest mean length fish is August, except for giant kōkopu (Panel A). The highest mean length for giant kōkopu is November. The month with heaviest mean fish for īnanga and kōaro is September. The month with heaviest mean banded kōkopu is October, and giant kōkopu is November (Panel B). The month with greatest mean depth īnanga and kōaro is September. The greatest mean depth banded kōkopu was in October, and for giant kōkopu was December (Panel C).

Table 27: Means and standard errors of continuous covariates. Standard error for date is supplied in days.

Species	n	Mean (standard error)					
		Length, mm	Weight, gm	Depth, mm	Latitude °	Longitude °	Date
Īnanga	12651	50.9 (0.26)	0.368 (0.008)	4.8 (0.04)	-41.4 (0.2)	173.2 (0.2)	6 Oct 2015 (3.2)
Kōaro	2295	52.1 (0.26)	0.531 (0.008)	5.5 (0.05)	-42.2 (0.3)	171.6 (0.2)	13 Oct 2015 (3.2)
Banded Kōkopu	2475	42.7 (0.26)	0.286 (0.008)	4.4 (0.05)	-40.5 (0.3)	173.2 (0.2)	18 Oct 2015 (3.2)
Giant Kōkopu	125	47.8 (0.26)	0.449 (0.008)	5.4 (0.05)	-41.8 (0.2)	171.9 (0.2)	21 Nov 2015 (3.1)

Multivariate Distributions

To test the assumption of normality for LDA Naïve Bayes I plotted multivariate normal quantile-quantile plots using R package mvtnorm (Mi, Miwa, & Hothorn, 2009). This checks multivariate normality by comparing the quantiles of Mahalanobis distances (the observed) to quantiles Chi squared distribution (the expected). Raw morphometric data was different from multivariate normal, except for the metrics of giant kōkopu (Figure 34). Allometric growth patterns usually mean that logging all length and weight will give a multivariate distribution that is closer to normal. Taking the log of this data did not bring the distribution closer to multivariate normal (Figures 33, 34). Juvenile fish approaching metamorphosis tend to shrink before reaching adulthood. Shrinking alters the allometric growth pattern. (Woods, 1968).

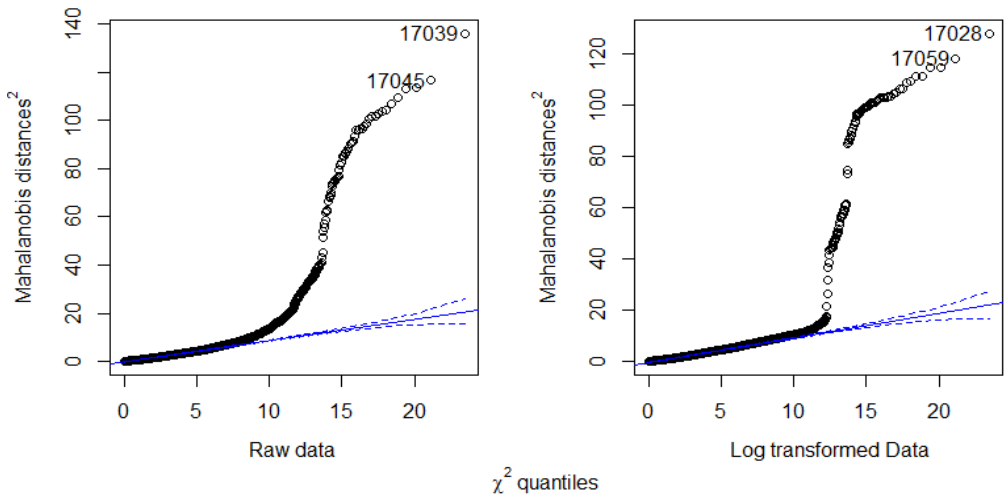


Figure 32: Natural and logged multivariate normality for all species from across the country for all morphological measures (weight, depth, and length). This data is far from multivariate normal (left panel). Even when the variables have all been logged, the data is still far from multivariate normal (right panel).

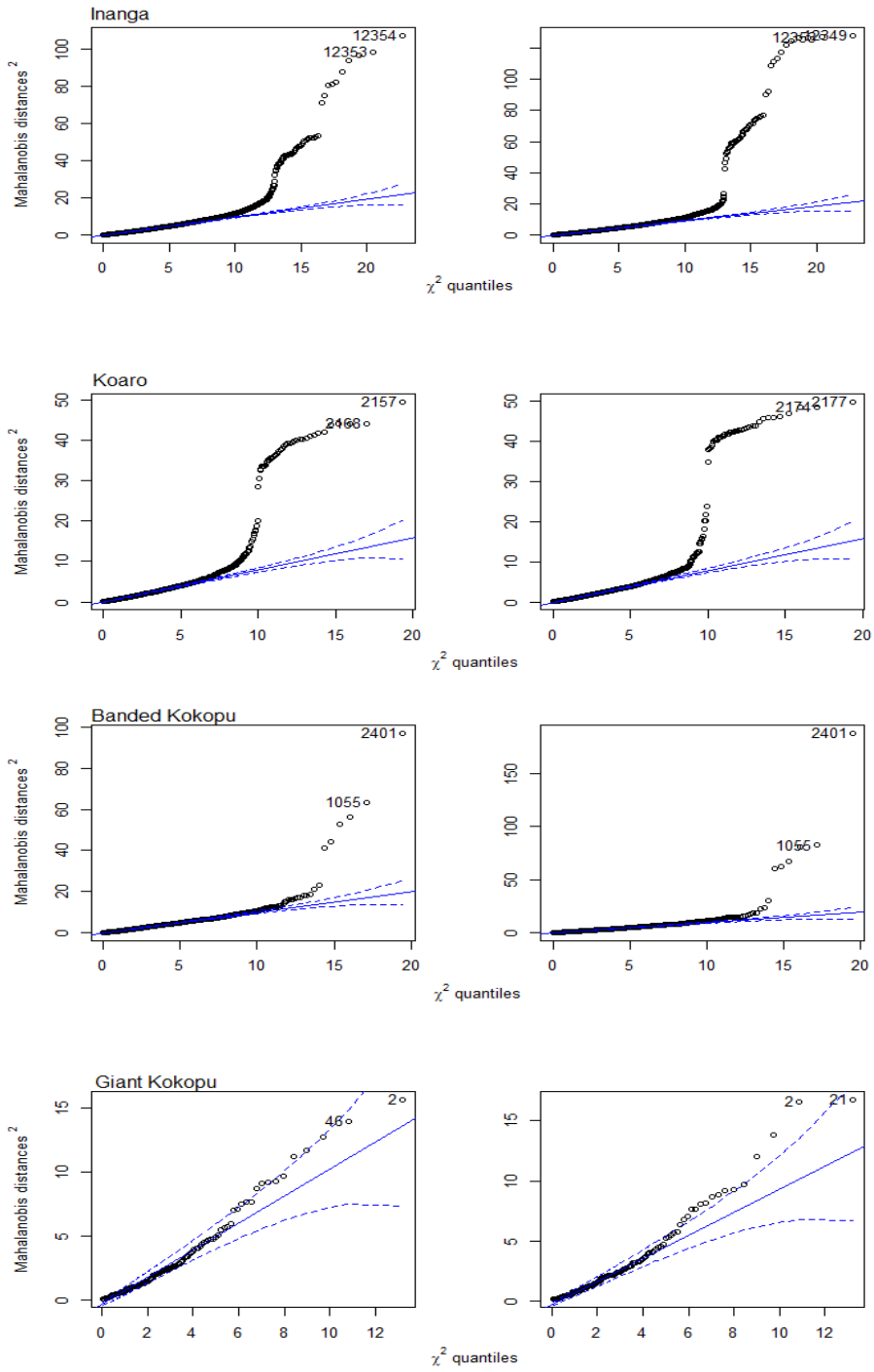


Figure 33: Natural and logged multivariate normality for each species to check the multivariate normality for weight length and depth. We expect that this would not be normal because fish have an allometric growth pattern. On the left are natural distributions. On the right are logged distributions. Giant kōkopu is the only measurements that are close to multivariate normal.

Section 4: Results

Model cross validation to choose the best model in each region

No species prevalence was adjusted for any models as the sampling was taken to be representative of the population in terms of species prevalence. The 10 fold cross validation error for each classification method differed across regions. At most it was 26.7% for Wellington data using Naïve Bayes (Table 26). The smallest CV error was 0% for Auckland data using LDA. In Waiarapa only īnanga were detected so there were no misclassifications.

Table 28: Correctly identified proportion of observations from 10-fold cross validation for all the country. Regions are listed from north to south. Country and region diversity indices are given.

Method	Species				Number correct	Number analysed	Total Proportion Correct	CV Error
	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga				
All regions	**Shannon Diversity = 0.822, Simpson Index = 0.56**							
MLR	0.2314	0.9759	0.6019	0.9702	14081	15403	0.9142	0.0858
LDA	0.1570	0.9899	0.5710	0.9655	13987	15403	0.9081	0.0919
Naïve Bayes	0.5200	0.8558	0.5747	0.8323	14031	17545	0.7997	0.2003
Decision Tree	0.0000	0.9071	0.4898	0.9723	15668	17545	0.8930	0.1070
Random Forest	0.5950	0.9873	0.7789	0.9833	14671	15403	0.9525	0.0475
Auckland	**Shannon Diversity = 0.822, Simpson Index = 0.65**							
MLR	-	0.8750	-	0.9750	85	88	0.9659	0.0341
LDA	-	1	-	1	88	88	1.0000	0.0000
Naïve Bayes	-	0.9737	-	0.9364	140	148	0.9459	0.0541
Decision Tree	-	0.9474	-	0.9636	142	148	0.9595	0.0405
Random Forest	-	1	-	0.9750	86	88	0.9773	0.0227
Waikato	**Shannon Diversity = 0.904, Simpson Index = 0.65**							
MLR	0.3333	0.9886	0.7143	0.9938	2347	2392	0.9812	0.0188
LDA	0	0.9981	0.6753	0.9921	2343	2392	0.9795	0.0205
Naïve Bayes	0.4444	0.8365	0.6667	0.9114	2713	3052	0.8889	0.1111
Decision Tree	0	0.9183	0.5387	0.9903	2932	3052	0.9607	0.0393
Random Forest	0.2222	0.9905	0.9464	0.9966	2347	2392	0.9812	0.0188
Bay of Plenty	**Shannon Diversity = 0.704, Simpson Index = 0.75**							
MLR	-	0.9795	0.6981	0.9923	1077	1103	0.9764	0.0236
LDA	-	0.9932	0.5094	0.9867	1064	1103	0.9646	0.0354
Naïve Bayes	-	0.9671	0.8704	0.9306	1364	1460	0.9342	0.0658
Decision Tree	-	0.8947	0.5185	0.9729	1384	1460	0.9479	0.0521
Random Forest	-	0.9589	0.6415	0.9934	1072	1103	0.9719	0.0281

Method	Species				Number correct	Number analysed	Total Proportion correct	CV error
	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga				
Taranaki	**Shannon Diversity = 1.273, Simpson Index = 0.45**							
MLR	-	0.9916	0.8000	0.9810	289	297	0.9731	0.0269
LDA	-	1	0.8500	0.9747	290	297	0.9764	0.0236
Naïve Bayes	-	0.9917	0.6500	0.8250	265	301	0.8804	0.1196
Decision Tree	-	0.9917	0.4500	0.9434	280	301	0.9302	0.0698
Random Forest	-	0.9916	0.3500	0.9937	282	297	0.9495	0.0505
Hawkes Bay	**Shannon Diversity = 0.642, Simpson Index = 0.78**							
MLR		0.9070	0.3962	0.9721	775	824	0.9405	0.0595
LDA	-	0.8837	0.3396	0.9753	766	824	0.9296	0.0704
Naïve Bayes	-	0.7917	0.434	0.9012	742	855	0.8678	0.1322
Decision Tree	-	0.7708	0.2075	0.9602	772	855	0.9029	0.0971
Random Forest	-	0.8286	0.4038	0.9830	687	735	0.9347	0.0653
Manawatu	**Shannon Diversity = 1.213, Simpson Index = 0.54**							
MLR	0.5000	1	0.6034	0.9782	776	838	0.9260	0.0740
LDA	0.3333	1	0.6293	0.9732	775	838	0.9248	0.0752
Naïve Bayes	0	0.9098	0.6923	0.8567	706	847	0.8335	0.1665
Decision Tree	0	0.9754	0.5128	0.9433	745	847	0.8796	0.1204
Random Forest	0.5000	1	0.6207	0.9832	781	838	0.9320	0.0680
Tasman	**Shannon Diversity = 1.488, Simpson Index = 0.39**							
MLR	0.2500	0.9877	0.7103	0.9187	1228	1388	0.8847	0.1153
LDA	0	0.9908	0.7259	0.9268	1239	1388	0.8927	0.1073
Naïve Bayes	0.5000	0.9189	0.6944	0.7760	1115	1411	0.7902	0.2098
Decision Tree	0	0.9459	0.6204	0.9013	1192	1411	0.8448	0.1552
Random Forest	0	0.9816	0.7850	0.9417	1266	1388	0.9121	0.0879
Wellington	**Shannon Diversity = 1.357, Simpson Index = 0.48**							
MLR	0	0.946667	0.410714	0.953488	340	396	0.8586	0.1414
LDA	0.5714	1	0.3571	0.9264	338	396	0.8535	0.1465
Naïve Bayes	0.7143	0.7	0.4035	0.8134	302	412	0.7330	0.2670
Decision Tree	0	0.9125	0.7193	0.9478	368	412	0.8932	0.1068
Random Forest	0	0.9867	0.6786	0.9612	360	396	0.9091	0.0909

Method	Species				Number correct	Number analysed	Total Proportion correct	CV Error
	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga				
Marlborough	**Shannon Diversity = 0.726, Simpson Index = 0.75**							
MLR	-	1	0.9316	0.9856	310	321	0.9657	0.0343
LDA	-	1	0.6842	0.9820	310	321	0.9657	0.0343
Naïve Bayes	-	0.9231	0.6000	0.9321	297	326	0.9110	0.0890
Decision Tree	-	0.9615	0.4000	0.9786	307	326	0.9417	0.0583
Random Forest	-	1	0.4211	1	310	321	0.9657	0.0348
Wairarapa	**Shannon Diversity = 0.000, Simpson Index = 1**							
MLR	-	-	-	1	40	40	1	0
LDA	-	-	-	1	40	40	1	0
Naïve Bayes	-	-	-	1	40	40	1	0
Decision Tree	-	-	-	1	40	40	1	0
Random Forest	-	-	-	1	40	40	1	0
Buller	**Shannon Diversity = 1.636, Simpson Index = 0.34**							
MLR	0.5294	0.9978	0.9119	0.9614	1744	1841	0.9473	0.0527
LDA	0.5588	1	0.8723	0.9601	1723	1841	0.9359	0.0641
Naïve Bayes	0.4706	0.9591	0.8448	0.8412	1722	1991	0.8649	0.1351
Decision Tree	0.4118	0.9836	0.8328	0.9837	1824	1991	0.9161	0.0839
Random Forest	0.5000	0.9932	0.9364	0.9686	1710	1789	0.9558	0.0442
Canterbury	**Shannon Diversity = 0.447, Simpson Index = 0.86**							
MLR		0.9423	0.6833	0.9977	1394	1419	0.9824	0.0176
LDA	-	1	0.6833	0.9962	1395	1419	0.9831	0.0169
Naïve Bayes	-	0.9630	0.7049	0.9790	1538	1589	0.9679	0.0321
Decision Tree	-	0.8333	0.6557	0.9946	1551	1589	0.9761	0.0239
Random Forest	-	0.9423	0.6667	0.9954	1390	1419	0.9796	0.0204
Westland	**Shannon Diversity = 1.335, Simpson Index = 0.49**							
MLR	0.8525	0.9821	0.7893	0.9547	2505	2726	0.9189	0.0811
LDA	0.8525	1	0.7542	0.9591	2497	2726	0.9160	0.0840
Naïve Bayes	0.8889	0.9589	0.5028	0.9103	2570	3124	0.8227	0.1773
Decision Tree	0.5397	0.9399	0.6459	0.9382	2700	3124	0.8643	0.1357
Random Forest	0.7869	0.9857	0.8428	0.9692	2560	2726	0.9391	0.0609

Method	Species				Number correct	Number analysed	Total Proportion correct	CV Error
	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga				
Otago	**Shannon Diversity = 0.835, Simpson Index = 0.696**							
MLR	-	0.9828	0.1616	0.9865	502	535	0.9383	0.0617
LDA	-	1	0.1612	0.9933	506	535	0.9458	0.0542
Naïve Bayes	-	0.9344	0.2571	0.9513	496	548	0.9051	0.0949
Decision Tree	-	0.9672	0.0857	0.9823	506	548	0.9234	0.0766
Random Forest	-	0.9728	0.0645	0.9910	501	535	0.9364	0.0636
Southland	**Shannon Diversity = 0.787, Simpson Index = 0.71**							
MLR	-	0.9649	0.3886	0.9522	1040	1195	0.8703	0.1297
LDA	-	1.0000	0.3829	0.9233	1042	1195	0.8720	0.1280
Naïve Bayes	-	0.9500	0.5801	0.9358	1285	1441	0.8917	0.1083
Decision Tree	-	0.8333	0.5745	0.9317	1308	1411	0.9270	0.0730
Random Forest	-	0.9474	0.6400	0.9709	1101	1195	0.9213	0.0787

Multinomial logistic regression

MLR was performed using R package nnet (Ripley, Venables, & Ripley, 2016). The CV error was 9% for national data. The lowest CV error rate was Canterbury with 1.76%. Banded kōkopu and īnanga were classified well with MLR (Tables 26, 27). The lowest classification rate of banded kōkopu using MLR was in Auckland (87.5%). Kōaro were often misclassified as īnanga using MLR, especially in Hawkes Bay, Otago and Southland where 56.6%, 83.87% and 61.14% of kōaro were misclassified as īnanga respectively (Table 27).

LDA

LDA was performed using R package MASS (Ripley et al., 2013) despite data not meeting the assumption of multivariate normality to see how LDA would perform. CV error was 9% for national data. Some regional data had low CV errors using LDA, especially Auckland (0%) and Taranaki (2%). All other regions and national data correctly classified over 95% of īnanga. Banded kōkopu were correctly classified most of the time with the lowest proportion correctly classified 88.37% in Hawkes Bay (Table 26). LDA did a poor job of classifying kōaro in all data sets (Tables 26, 28). Using LDA, giant kōkopu were most often misclassified as banded kōkopu in national data, and Waikato. At worst, LDA classified 83.87% of kōaro as īnanga in Otago. The highest proportion of correctly classified kōaro was 85.00% in Taranaki and 87.23% in Buller. Īnanga from the Otago data had a correct classification rate of ~95% (Table 28).

Naïve Bayes

Naïve Bayes was performed using R package e1071 (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2017). Gaussian Bayes was used despite the species data not meeting multivariate normality (Figure 33) to see how the model would perform. No discretisation was used as finding the best discretisation level is problematic. Few regions had high classification rates with Naïve Bayes (Tables 26, 29).

In the national data giant kōkopu were most often misclassified as īnanga (20%) (Table 29). With data from the Tasman region, 50% of giant kōkopu were misclassified as kōaro. Banded kōkopu were classified correctly above 80% of the time, except in the Hawkes Bay (77.08%) and Wellington (70%) regions (Table 29). From Wellington data, banded kōkopu were misclassified as giant kōkopu 28.75% (Table 29). Kōaro were poorly classified by Naïve Bayes. The highest rate of correct classification for kōaro was 87.04% in the Bay of Plenty region (Table 3). In all regions, kōaro were most often misclassified as īnanga. 71% of kōaro were misclassified as īnanga in Otago data (Table 29).

Decision Tree

Decision tree was performed using R package rpart (Therneau & Atkinson, 2018). Default settings were used. Trees were not pruned. All trees were grown to have terminal nodes that had a complexity parameter of 0.01.

There were overall poor classification rates using decision tree (Tables 26, 30). The highest correct classification rate was for īnanga in Canterbury (Table 26). Īnanga were classified with greater than 90% accuracy in each region (Table 26). Kōaro were frequently misclassified as īnanga (Table 30). 73% of kōaro were misclassified as īnanga from Hawkes Bay data. 91% of Otago kōaro were misclassified as īnanga. The highest rate of correctly classified kōaro using decision trees was 83.28% from Buller (Table 26). Banded kōkopu were correctly classified with 90% and above accuracy for all regions and for national data. Giant kōkopu were unable to be distinguished from the national data; they were frequently misclassified as either kōaro (44%) or īnanga (38%)(Table 30). Buller and Westland, had giant kōkopu correctly classified with 41.18% for Buller, and 53.97% for Westland (Table 30).

Random Forest

Random Forest was performed using R package randomForest (Liaw & Wiener, 2002). Overall, random forest had the lowest CV error, but did not distinguish kōaro from īnanga well for Otago or Marlborough data (Table31). The lowest proportion of correctly classified kōaro came from Otago with 94% misclassified as īnanga. Banded kōkopu were classified with above 80% purity for all data (Table 31). Banded kōkopu were 100% correctly classified in the Auckland region, but īnanga were not (Table 26). In the regionally pooled data, there were 60% correctly classified giant kōkopu, with 25% misclassified as kōaro. In regional data giant kōkopu were classified incorrectly more than 50% (Table 26). Tasman and Wellington region giant kōkopu misclassified as īnanga, kōaro, or banded kōkopu (Table 31).

By Species

Giant kōkopu: sample prevalence, 1%

Where giant kōkopu were detected, the best classifications were made using Naïve Bayes in Westland (88.89%) and Wellington (71.43%). The worst detection rate of giant kōkopu was 0% in Manawatu. Using national data Random Forest performed the best with 59.5% correctly classified observations (Table 31). There was no one species that giant kōkopu was most often misclassified as (Tables 27, 28, 29, 30, 31).

Banded kōkopu: sample prevalence, 14%

LDA was 100% accurate at classifying banded kōkopu in Auckland, Canterbury, Marlborough, Otago, Southland, Taranaki, Westland, Wellington, Buller, and Manawatu (Table 26). The lowest rate of correctly classified banded kōkopu using LDA was 88.37% in Hawkes Bay (Table 26). Multinomial Logistic regression was 100% accurate for banded kōkopu in Marlborough and Manawatu. Random Forest was 100% accurate for banded kōkopu in Auckland, Marlborough, and Manawatu. The lowest classification rate was from Wellington data using Naïve Bayes. Using national data, the best method was LDA with 98.99% of Banded kōkopu correctly classified (Table 3). When banded kōkopu were misclassified it was most often as īnanga (Tables 27, 28, 29, 30, 31).

Kōaro: sample prevalence, 13%

Kōaro were not classified to 100% accuracy using any method (Table 26). The highest correct classification rate was 94.64% using Random forest on Buller data. The best method for national data was random forest with a correct classification rate of 77.89%. Otago had the lowest rate of correctly classified kōaro regardless of the classification method (Table 26). In Otago īnanga and kōaro have similar morphological characteristics (Figures 27, 28, 29). Each classification method did a poor job of distinguishing kōaro in Otago data (Table 26). When they were misclassified, kōaro were mist often misclassified as īnanga (Tables 27, 28, 29, 30, 31).

Īnanga: sample prevalence, 72%

There were only īnanga detected in Wairarapa, so all classification accuracies were 100% (Table 26). Īnanga were classified correctly 100% of the time using LDA in Auckland and using Random Forest in Marlborough. The lowest classification rate for īnanga from the regions data was Tasman for all methods except LDA. From national data the method with the highest correct classification rate was Random Forest (Table 26). When they were misclassified, īnanga were most often misclassified as kōaro (Tables 27, 28, 29, 30, 31).

Table 29: Selected confusion matrices for multinomial logistic regression. All errors are CV error. Kōaro were frequently misclassified as īnanga. Īnanga had the highest classification rate

		Observed							
		Count				Percentage			
MLR	Predicted	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga
National	Giant Kōkopu	28	6	10	0	23.14	0.26	0.47	0.00
	Banded Kōkopu	20	2225	25	62	16.53	97.59	1.17	0.57
	Kōaro	68	17	1285	262	56.20	0.75	60.19	2.41
	Īnanga	5	32	815	10543	4.13	1.40	38.17	97.02
Hawkes Bay	Banded Kōkopu		39	2	2		90.70	3.77	0.27
	Kōaro		0	21	11		0	39.62	1.51
	Īnanga		4	30	715		9.30	56.60	98.21
Otago	Banded Kōkopu		57	0	1		98.28	0	0.22
	Kōaro		0	5	5		0	16.13	1.12
	Īnanga		1	26	440		1.72	83.87	98.65
Southland	Banded Kōkopu		55	0	1		96.49	0	0.10
	Kōaro		1	68	45		1.75	38.86	4.67
	Īnanga		1	107	917		1.75	61.14	95.22

Table 30: selected confusion matrices for linear discriminant analysis. Percentage errors are CV error. Kōaro were frequently misclassified as īnanga, especially in Otago where 83.87% of kōaro were misclassified as īnanga. In Otago Banded kōkopu are classified with 100% accuracy.

		Observed							
		Count				Percentage			
LDA	Predicted	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga
National	Giant Kōkopu	19	2	3	1	15.70	0.09	0.14	0.01
	Banded Kōkopu	31	2257	35	125	25.62	98.99	1.64	1.15
	Kōaro	63	4	1219	249	52.07	0.18	57.10	2.29
	Īnanga	8	17	878	10492	6.61	0.75	41.12	96.55
Waikato	Giant Kōkopu	0	0	2	0	0.00	0.00	2.60	0.00
	Banded Kōkopu	7	524	3	14	77.78	99.81	3.90	0.79
	Kōaro	1	0	49	4	11.11	0.00	63.64	0.22
	Īnanga	1	1	23	1763	11.11	0.19	29.87	98.99
Otago	Banded Kōkopu		58	0	1		100.00	0.00	0.22
	Kōaro		0	5	2		0.00	16.13	0.45
	Īnanga		0	26	443		0.00	83.87	99.33

Table 31: selected confusion matrices for Naïve Bayes. Percentage errors are CV error. Giant kōkopu from national data were frequently misclassified as īnanga or kōaro. In the Tasman region, giant kōkopu were misclassified as kōaro. Banded kōkopu had a relatively low correct classification rate of 70%, 28.75% of banded kōkopu were misclassified as giant kōkopu. 71.43% of kōaro were misclassified as īnanga in Otago.

		Observed							
		Count				Percentage			
Naïve Bayes	Predicted	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga
National	Giant Kōkopu	65	10	0	8	52.00	0.40	0.00	0.06
	Banded Kōkopu	11	2118	11	802	8.80	85.58	0.48	6.34
	Kōaro	24	2	1319	1311	19.20	0.08	57.47	10.36
	Īnanga	25	345	965	10529	20.00	13.94	42.05	83.23
Wellington	Giant Kōkopu	5	23	11	35	71.43	28.75	19.3	13.06
	Banded Kōkopu	1	56	0	3	14.29	70.00	0.00	1.12
	Kōaro	1	0	23	12	14.29	0.00	40.35	4.48
	Īnanga	0	1	23	218		1.25	40.35	81.34
Tasman	Giant Kōkopu	2	2	4	2	50.00	0.60	1.23	0.27
	Banded Kōkopu	0	306	3	35	0.00	91.89	0.93	4.67
	Kōaro	2	0	225	131	50.00	0.00	69.44	17.47
	Īnanga	0	25	92	582	0.00	7.51	28.40	77.60
Otago	Banded Kōkopu		57	1	6		93.44	2.86	1.33
	Kōaro		0	9	16		0.00	25.71	3.54
	Īnanga		4	25	430		6.56	71.43	95.13

Table 32: Selected confusion matrices for Decision Tree. Percentage errors are CV error. Giant Kōkopu were unable to be identified in national data. In Hawkes Bay, giant kōkopu were frequently misclassified as īnanga. 91.43% of kōaro from Otago were misclassified as īnanga.

		Observed							
		Count				Percentage			
Decision Tree	Predicted	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga
National	Giant Kōkopu	0	0	0	0	0.00	0.00	0.00	0.00
	Banded Kōkopu	22	2245	15	149	17.60	90.71	0.65	1.18
	Kōaro	55	12	1124	202	44.00	0.48	48.98	1.60
	Īnanga	48	218	1156	12299	38.40	8.81	50.37	97.23
Hawkes Bay	Banded Kōkopu		37	0	16		77.08	0.00	2.12
	Kōaro		0	11	14		0.00	20.75	1.86
	Īnanga		11	42	724		22.92	79.25	96.02
Otago	Banded Kōkopu		59	0	3		96.72	0.00	0.66
	Kōaro		0	3	5		0.00	8.57	1.11
	Inānga		2	32	444		3.28	91.43	98.23

Table 33: Selected confusion matrices for Random Forest. All percentages are CV error. In national data, giant kōkopu were misclassified 24.79% of the time as kōaro. In Marlborough, kōaro were frequently misclassified as īnanga. In Otago 93.55% of kōaro were misclassified as īnanga. Giant kōkopu in Tasman and Wellington were all misclassified. In Tasman giant kōkopu were misclassified as kōaro, and in Wellington giant kōkopu were either misclassified as banded kōkopu or kōaro.

		Observed							
		Count				Percentage			
Random Forest	Predicted	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga	Giant Kōkopu	Banded Kōkopu	Kōaro	Īnanga
All Regions	Giant Kōkopu	72	0	3	0	59.50	0.00	0.14	0.00
	Banded Kōkopu	9	2251	11	21	7.44	98.73	0.52	0.19
	Kōaro	30	8	1663	161	24.79	0.35	77.89	1.48
	Īnanga	10	21	458	10685	8.26	0.92	21.45	98.33
Marlborough	Banded Kōkopu		24	0	0		100.00	0.00	0.00
	Kōaro		0	8	0		0.00	42.11	0.00
	Īnanga		0	11	278		0.00	57.89	100
Wellington	Giant Kōkopu	0	1	2	0	0.00	1.33	3.57	0.00
	Banded Kōkopu	4	74	0	2	57.14	98.67	0.00	0.78
	Kōaro	3	0	38	8	42.86	0.00	67.86	3.10
	Īnanga	0	0	16	248	0.00	0.00	28.57	96.12
Tasman	Giant Kōkopu	0	0	0	0	0.00	0.00	0.00	0.00
	Banded Kōkopu	0	319	3	4	0.00	98.15	0.93	0.54
	Kōaro	4	2	252	39	100.00	0.62	78.50	5.28
	Īnanga	0	4	66	695	0.00	1.23	20.56	94.17
Otago	Banded Kōkopu		57	0	2		98.28	0.00	0.45
	Kōaro		0	2	2		0.00	6.45	0.45
	Īnanga		1	29	442		1.72	93.55	99.10

Section 5: Discussion

General Discussion

Six classification methods, Multinomial logistic regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naïve Bayes, Decision Tree, and Random Forest were applied to the data to classify observations into the four species īnanga (*Galaxias maculatus*), kōaro (*Galaxias brevipennis*), banded kōkopu (*Galaxias fasciatus*), and giant kōkopu (*Galaxias argentus*). The best method, found using 10 fold cross validation (CV), was Random Forest (Table 28). The best method for distinguishing kōaro was Random Forest, but it still misclassified 30% of kōaro (Table 33). In national data, Naïve Bayes identified kōaro correctly ~60% of the time (Table 31). The best method for distinguishing giant kōkopu correctly was Naïve Bayes which identified with 52% accuracy in national data and at best 89% accuracy in Westland data (Table 31). The model that performed the best by CV error overall was Random Forest, however, Random Forest did not do well distinguishing giant kōkopu or kōaro (Table 33).

QDA was abandoned because two single species (giant kōkopu and kōaro) variance-covariance matrices could not to be estimated. There were too few data for giant kōkopu and kōaro in some regions (Appendix 1.1). Getting more data for those species would mean that the variance-covariance matrices could be estimated, but, QDA is not able to distinguish species that are difficult to separate using some linear combination of the covariates (Figure 8). Īnanga and kōaro had similar morphometrics (Figure 29) so I do not think that QDA would offer any improvement over the classification rates presented here.

Gaussian Naïve Bayes assumes multivariate normally distributed data (Hastie & Tibshirani, 2009). The data for each species presented here was not multivariate normal (Figure 32), and Naïve Bayes performed poorly overall (Table 28). Discretisation was not tried as finding the 'right' level of discretisation is difficult to defend. Multivariate normal data was also an assumption for LDA, but LDA classified banded kōkopu and īnanga well (Table 30). Banded kōkopu were easily distinguished across all methods (Table 28). This suggests that banded kōkopu are easy to classify.

How the prevalence of each species changes across the season between regions is not well understood (Goodman, 2018). We assumed that the samples were representative of the populations in terms of morphometric measures and in terms of species prevalence through the season. However, this may not the best assumption as the sampling effort was not recorded and may not have been consistent across

the season, or across the country (Figure 25) (Table A1). Morphological measures of each species varied across the season (Figure 31). Although, this variation may not have been sufficient to make classifications easier across the season as the ranges of morphometrics for each species were not distinct (Figure 31).

There were differences in species prevalence between the regions. Not all species were detected in all regions (Figure 25). Regions with the lowest number of species were classified the best, such as Wairarapa, and Auckland (Table 28). The prevalence of īnanga is highest (Figure 26). For the whole country, if we took every observation and classify it as īnanga, then the classification rate would be 72%. No matter the method, īnanga were almost always identified with the highest accuracy (Table 28). If we can positively identify īnanga near 100% of the time and remove them from the classification, then the job is to distinguish kōaro and giant kōkopu as banded kōkopu are easily identified.

Let us compare the best method Random Forest to the most intuitive method, Decision Tree as Decision Trees are similar to biological keys used for species identification. With Decision Tree classification it is possible to print off the decision tree and classify fish without a computer(Figure 14); whereas Random Forest requires a computer as the classification process is too complex (Figure 21). The national correct classification rate for Random Forest for īnanga was 98.33%, and 77.89% for kōaro. For Decision Tree, the correct classification rate for īnanga was 97.23% and 48.98% for kōaro. Additionally, Decision Tree classified giant kōkopu correctly 0% of the time (Table 28). Having an intuitive method such as decision tree is helpful to understand how the classification is being made, but when the benefit is correctly classifying more of a rare species, perhaps inputting measurements into a computer is not such a cost.

Although Random Forest made the best predictions overall for almost every region and nationally, in cross validation, 21.45% of kōaro were misclassified as īnanga (Table 33). In general, īnanga and kōaro were the most difficult to distinguish for all methods (Tables 29, 30, 31, 32, 33). The density of morphometric measures for īnanga and kōaro are almost indistinguishable (Figure 30). In particular, the morphometrics of Otago īnanga and kōaro are almost identical (Figure 32). Obtaining more observations of kōaro may help correctly classify more kōaro, but given how mixed the distributions of īnanga and kōaro are, adding more data is unlikely to help. In microscope comparisons, the major discerning feature between īnanga and kōaro is the offset of the dorsal and anal fin (Figure 24) (Yungnickel 2017). The ‘depth’ measure was an attempt to capture the difference in offset of pectoral and anal fins between

īnanga and kōaro. In future work, assessing variable importance to classification would verify if this measure is capturing the differences in this feature by species.

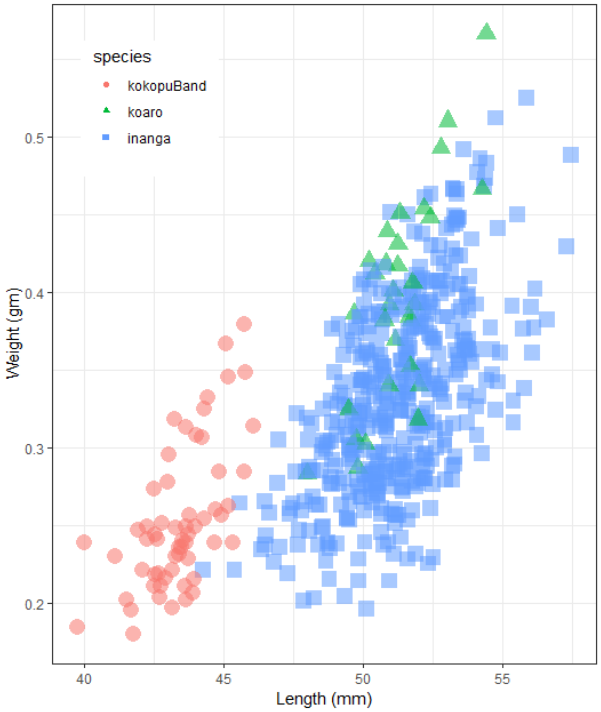


Figure 32: Plot of Length vs Weight for Otago data. Īnanga and kōaro are virtually indistinguishable using these two metrics, but banded kōkopu is easily distinguished (red). This explains in part why each of the methods had low correct classification rates for kōaro, why kōaro were most often misclassified as īnanga, and why banded kōkopu were easy to correctly classify in Otago.

If any of these models were to be applied to data from other years' data, or data from other New Zealand rivers that were not sampled for this project, the correct classification rate would probably be similar for all species. Collecting more data in years to come would mean we could compare model errors between the years. More data may not offer more power to distinguish between kōaro and īnanga with the current measures. Applying these methods to other fisheries would yield better results as most fisheries rely on mature fish which have more defined morphological characteristics to distinguish species.

Other supervised classification methods that could be used to classify whitebait into species include Support Vector Machines, Neural Networks, or Ensemble Learning (Hastie & Tibshirani, 2009). For this dataset, these methods would not yield useful results as they require vast quantities of data (Hastie & Tibshirani, 2009).

Future work

Future work could involve creating a classification model that incorporates the nestedness of rivers within regions, and assessing the spatial distribution. There is evidence to suggest that whitebait have a metapopulation structure (Egan, 2017) and incorporating nested spatial data would capture some of this variability. Jones and Checkley classified fish from otoliths (2017). Otoliths have been explored for whitebait population dynamics (Egan 2017) but not for species classification.

The sampling method that was used to collect this data meant that there are other features of the fish characterised that were not able to be characterised. For example, kōaro are known to climb up the bucket when they are sampled. In sampling, the buckets were mixed because kōaro tend to be near the top (Yungnickel 2017) and we know that kōaro are difficult to identify. Stratified sampling of buckets might give higher proportions of kōaro to measure. We could then compare stratified samples from rivers where fewer species have been detected to rivers where more species, in particular kōaro and banded kōkopu, have been detected. Or record if a fish was climbing out of a bucket at the time of sampling.

Supervised learning algorithms have been used to classify whitebait with varying success. At best, Random Forest correctly classified ~95% of observations by finding differences between morphometrics in whitebait species. However, kōaro and giant kōkopu are still difficult to distinguish using Random Forest. These classification methods use measurements that are similar to measures that are compared for microscope species identification. For these methods to increase the rate of correct classifications between similar species, other variables will need to be characterised, for example if a fish climbs up the bucket. Capturing this characteristic is likely to greatly improve supervised classification. Identifying species is vital to characterising the biodiversity of the fishery. Understanding how the species composition changes over time will contribute to a better understanding of sustainability of the fishery. Our study contributes a step towards a fast and cheap way of identifying galaxiid whitebait into species based on their morphological characteristics.

Bibliography

Agresti, A. (2013). Categorical Data Analysis (3rd ed.). New Jersey: John Wiley & Sons.

Allibone, R. M., & Caskey, D. (2000). Timing and habitat of koaro (Galaxias brevipinnis) spawning in streams draining Mt Taranaki, New Zealand. New Zealand Journal of Marine and Freshwater Research, 34(4), 593-595.

Allibone, R., David, B., Hitchmough, R., Jellyman, D., Ling, N., Ravenscroft, P., & Waters, J. (2010). Conservation status of New Zealand freshwater fish, 2009. New Zealand Journal of Marine and Freshwater Research, 44(4), 271-287. doi:10.1080/00288330.2010.514346

Baker, C. F., Egan, E. M. C., & Gee, E. (2018). Potential options for regulation changes to the NZ whitebait fishery (2018160HN). Retrieved from Hamilton:

Bonnett, M., McDowall, R. M., & Sykes, J. (2002). Critical habitats for the conservation of giant kokopu, Galaxias argenteus (Gmelin, 1789): Department of Conservation Wellington.

Breiman, L. (2001). Random forests. 45(1), 5-32.

Charteris, S. C., & Ritchie, P. A. (2002). Identification of galaxiid nests, emigrating larvae and whitebait, using mitochondrial DNA control region sequences. New Zealand Journal of Marine and Freshwater Research, 36(4), 789-795.

D'Elia, M., Patti, B., Bonanno, A., Fontana, I., Giacalone, G., Basilone, G., & Fernandes, P. J. F. R. (2014). Analysis of backscatter properties and application of classification procedures for the identification of small pelagic fish species in the Central Mediterranean. 149, 33-42.

Dijkstra, L. H., & McDowall, R. M. (1997). Electrophoretic identification of whitebait species. Conservation Advisory Science Notes, 153, 1-13.

Dunn, N. R., Allibone, R. M., Closs, G. P., Crow, S. K., David, B. O., Goodman, J. M., Griffiths, M., Jacks, D. C., Ling, N., Waters, J. M., & Rolfe, J.R. (2018). Conservation status of New Zealand freshwater fishes, 2017: Publishing Team, Department of Conversation.

Egan, E. M. C. (2017). Early life history of the amphidromous galaxiid īnanga: disentangling the consequences for their migratory dynamics, population stucture and adult growth. (PhD PhD thesis), University of Canterbury, Christchurch.

Elmqvist, T., Folke, C., Nyström, M., Peterson, G., Bengtsson, J., Walker, B., & Norberg, J. (2003). Response diversity, ecosystem change, and resilience. *Frontiers in Ecology and the Environment*, *1*(9), 488-494.

Friedman, J. H., Kohavi, R., & Yun, Y. (1996, August). Lazy decision trees. In AAAI/IAAI, Vol. 1 (pp. 717-724).

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Gaston, K. J., & O'Neill, M. A. (2004). Automated species identification: why not? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1444), 655-667.

Glova, G. J. (2003). A test for interaction between brown trout (Salmo trutta) and inanga (Galaxias maculatus) in an artificial stream. Ecology of Freshwater Fish, 12(4), 247-253. doi:10.1046/j.1600-0633.2003.00019.x

Goodman, J. M., Dunn, N. R., Ravenscroft, P. J., Allibone, R. M., Boubée, J. A. T., David, B. O., . . . Rolfe, J. R. (2014). Conservation status of New Zealand freshwater fish, 2013. New Zealand Threat Classification Series, 7, 12p.

Goodman, J. (2018). Conservation, Ecology and Management of Migratory Galaxiids and the Whitebait Fishery: A Summary of Current Knowledge and Information Gaps. *Wellington (New Zealand): Department of Conservation (DoC)*.

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Peito-Piraquive, E., Jenieiro E., … Patti, C. (2010). IPez: an expert system for the taxonomic identification of fishes based on machine learning techniques. Fisheries Research, 102(3), 240-247.

Hickford, M. J. H., & Schiel, D. R. (2010). Eggs are being laid: experimental rehabilitation of the riparian spawning habitat of a diadromous fish, *Galaxias maculatus*. Paper presented at the New Zealand Marine Sciences Society Conference, Wellington.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112): Springer.

Jones, W. A., & Checkley Jr, D. M. (2017). Classification of otoliths of fishes common in the Santa Barbara Basin based on morphology and chemical composition. Canadian Journal of Fisheries and Aquatic Sciences, 74(8), 1195-1207.

Kahle, D., & Wickham, H. J. T. R. j. (2013). ggmap: Spatial Visualization with ggplot2. 5(1), 144-161.

Keylock, C. (2005). Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. Oikos, 109(1), 203-207.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at the Ijcai.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. Editorial Boards, Publishing Council, 31, 249-268.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News: R News. Retrieved from https://CRAN.R-project.org/doc/Rnews/

McDowall, R. M. (1964). A consideration of the question "What are whitebait?". Tuatara, 12(3), 134-146.

McDowall, R. M. (1965). The composition of the New Zealand whitebait catch, 1964. New Zealand Journal of Science, 8(3), 285-300.

McDowall, R. M. (1970). The galaxiid fishes of New Zealand. Bulletin of the Museum of Comparative Zoology at Harvard College, 139(7), 341-431.

McDowall, R. M. (1996). Managing the New Zealand whitebait fishery: a critical review of the role and performance of the Department of Conservation. NIWA Science and Technology Series, 32, 1-39.

McLean, F., Barbee, N. C., & Swearer, S. E. (2007). Avoidance of native versus non-native predator odours by migrating whitebait and juveniles of the common galaxiid, Galaxias maculatus.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability

Mi, X., Miwa, T., & Hothorn, T. J. R. J., Nr. 1. (2009). mvtnorm: New numerical algorithm for multivariate normal probabilities. 1(1), 37-39.

Milborrow, S. M. (2018). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. Retrieved from https://CRAN.R-project.org/package=rpart.plot

Orchard, D. S. E., Hickford, M. J. H., & Schiel, D. R. (2018). Earthquake-induced habitat migration in a riparian spawning fish has implications for conservation management. Aquatic Conservation: Marine and Freshwater Ecosystems, 28(3), 702-712. doi:doi:10.1002/aqc.2898

Remane, A., & Schlieper, C. (1972). Biology of brackish water.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package ‘mass’. Cran R.

Ripley, B., Venables, W., & Ripley, M. B. (2016). Package ‘nnet’. R package version, 7, 3-12.

Rowe, D. K., & Kelly, G. (2009). Duration of the oceanic phase for inanga whitebait (Galaxiidae) is inversely related to growth rate at sea. In A. Haro, K. L. Smith, R. A. Rulifson, C. M. Moffitt, R. J. Klauda, M. J. Dadswell, R. A. Cunjak, J. E. Cooper, K. L. Beal, & T. S. Avery (Eds.), Challenges for diadromous fishes in a dynamic global environment (pp. 343-354). Halifax: American Fisheries Society.

Taylor, M. J. (2002). The national inanga spawning database: trends and implications for spawning site management. Science for Conservation, 188, 1-37.

Therneau, T., & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees. Retrieved from https://CRAN.R-project.org/package=rpart

Whitebait Fishing Regulations, 1994 (New Zealand)

Woods, C. S. (1968). Growth characteristics, pigmentation and the identification of whitebait (Galaxias spp., Salmonoidea). New Zealand Journal of Marine and Freshwater Research, 2(2), 162-182.

Yungnickel, M. R. (2017). New Zealand's whitebait fishery: Spatial and temporal variation in species composition and morphology.

Appendix 1: Region Descriptions

Table A1: Regions and river information. Each river is allocated to one region. The largest number of rivers is one region is 15. The Mokau river in the Waikato was sampled on the most occasions. At most there were four samplers working on a river. There were two rivers that had four samplers working; they were the Buller river in Buller, and the Avon river in Canterbury.

region	river	n	Latitude	Longitude	Earliest date	Latest date	Sampling occasions	Samplers
Auckland	Hoteo	148	-36.4	174.5	15 October 2015	19 November 2015	3	2
BOP	Whakatane	285	-37.9	177.0	23 July 2015	5 October 2015	6	3
BOP	Rangitaiki	132	-37.9	176.9	30 August 2015	6 October 2015	3	2
BOP	Kaituna	572	-37.8	176.4	6 July 2015	18 November 2015	9	1
BOP	Nukuhou	138	-38.0	177.1	22 July 2015	11 September 2015	5	1
BOP	Whangaparaoa	102	-37.6	178.0	20 August 2015	1 September 2015	2	1
BOP	Otara	43	-38.0	176.9	6 October 2015	6 October 2015	1	1
BOP	Tarawera	40	-37.9	176.8	7 September 2015	7 September 2015	1	1
BOP	Tuapiro	81	-37.5	175.9	8 October 2015	8 October 2015	1	1
BOP	Waiaua	65	-38.0	176.9	15 October 2015	15 October 2015	1	1
BOP	Waiotahi	2	-38.0	177.2	12 October 2015	12 October 2015	1	1
Buller	Buller	511	-41.8	171.6	22 July 2015	18 November 2015	7	4
Buller	Mokihinui	384	-41.5	171.9	22 July 2015	17 November 2015	5	3
Buller	Punakaiki	223	-42.1	171.3	14 September 2015	15 December 2015	5	3
Buller	Orowaiti	170	-41.8	171.6	15 September 2015	9 November 2015	3	3
Buller	Karamea	465	-41.3	172.1	15 September 2015	14 November 2015	4	2
Buller	Oparara	137	-41.2	172.1	25 September 2015	11 October 2015	2	1
Buller	LilWanganui	57	-41.4	172.1	24 September 2015	24 September 2015	1	1
Buller	Okari	44	-41.8	171.5	19 September 2015	19 September 2015	1	1
Canterbury	Avon	424	-43.5	172.7	4 August 2015	21 December 2015	7	4
Canterbury	Waimakariri	293	-43.4	172.7	12 September 2015	20 December 2015	7	1
Canterbury	Saltwater	244	-43.3	172.7	28 August 2015	23 December 2015	6	1
Canterbury	Ashley	162	-43.3	172.7	30 September 2015	26 November 2015	4	1
Canterbury	Hapuku	47	-42.3	173.7	30 November 2015	30 November 2015	1	1
Canterbury	Heathcote	43	-43.6	172.7	24 November 2015	24 November 2015	1	1
Canterbury	Kowai	50	-42.4	173.6	21 November 2015	21 November 2015	1	1
Canterbury	Lyell	46	-42.4	173.7	18 November 2015	18 November 2015	1	1
Canterbury	Okains	40	-43.7	173.1	24 October 2015	24 October 2015	1	1

Canterbury	Opihi	40	-44.3	171.3	2 November 2015	2 November 2015	1	1
Canterbury	Orari	40	-44.2	171.4	10 November 2015	10 November 2015	1	1
Canterbury	Pawsons	40	-43.8	172.9	21 October 2015	21 October 2015	1	1
Canterbury	RobinsonsBay	40	-43.8	173.0	25 October 2015	25 October 2015	1	1
Canterbury	Styx	40	-43.4	172.7	27 August 2015	27 August 2015	1	1
Canterbury	Waihao	40	-44.8	171.2	29 November 2015	29 November 2015	1	1
HawkesBay	Ngaruroro	212	-39.6	176.9	24 July 2015	7 October 2015	5	3
HawkesBay	Tukituki	109	-39.6	176.9	3 July 2015	4 September 2015	4	2
HawkesBay	Wairoa	196	-39.1	177.3	5 October 2015	10 November 2015	4	2
HawkesBay	Tutaekuri	202	-39.6	176.9	8 October 2015	13 November 2015	4	1
HawkesBay	Porangahau	95	-40.3	176.6	8 October 2015	27 October 2015	2	1
HawkesBay	Clive	41	-39.6	176.9	9 October 2015	9 October 2015	1	1
Manawatu	Rangitikei	507	-40.3	175.2	9 September 2015	19 November 2015	7	2
Manawatu	KaiIwi	82	-39.9	174.9	28 September 2015	30 November 2015	2	1
Manawatu	Whangaehu	166	-40.0	175.1	7 October 2015	18 October 2015	2	1
Manawatu	Manawatu	40	-40.5	175.2	14 September 2015	14 September 2015	1	1
Manawatu	Owahanga	52	-40.6	176.3	17 November 2015	17 November 2015	1	1
Marlborough	WairauDiv	187	-41.4	174.0	3 September 2015	28 November 2015	4	2
Marlborough	Awatere	42	-41.6	174.2	30 November 2015	30 November 2015	1	1
Marlborough	Opawa	55	-40.8	174.0	24 October 2015	24 October 2015	1	1
Marlborough	Wairau	42	-41.5	174.1	13 October 2015	13 October 2015	1	1
Otago	Taeri	226	-46.1	170.2	29 October 2015	20 November 2015	4	2
Otago	Waitaki	134	-44.9	171.1	28 October 2015	22 November 2015	3	2
Otago	Kakanui	85	-45.2	170.9	28 October 2015	30 November 2015	2	1
Otago	Owaka	63	-46.4	169.7	30 October 2015	28 November 2015	2	1
Otago	Shag	40	-45.5	170.8	28 October 2015	28 October 2015	1	1
Southland	Titiroa	216	-46.6	168.8	17 August 2015	11 November 2015	7	2
Southland	Mataura	225	-46.6	168.7	16 August 2015	30 November 2015	4	2
Southland	Aparima	325	-46.3	168.0	7 August 2015	21 December 2015	9	1
Southland	Waiau	498	-46.2	167.6	2 August 2015	26 November 2015	9	1
uthland	Oreti	120	-46.5	168.7	10 August 2015	28 September 2015	3	1
Southland	Waikawa	57	-46.6	169.1	11 November 2015	11 November 2015	1	1
Taranaki	Onaero	161	-39.0	174.4	10 September 2015	24 September 2015	2	2

Taranaki	Waingongoro	51	-39.6	174.2	27 September 2015	27 September 2015	1	1
Taranaki	Waitara	89	-39.0	174.2	22 September 2015	22 September 2015	1	1
Tasman	Aorere	138	-40.7	172.7	7 July 2015	13 October 2015	3	2
Tasman	Takaka	902	-40.8	172.8	5 July 2015	17 December 2015	12	1
Tasman	Wainui	309	-40.8	172.9	16 September 2015	16 December 2015	5	1
Tasman	Motueka	40	-41.1	173.0	31 July 2015	31 July 2015	1	1
Tasman	Parapara	22	-40.7	172.7	5 September 2015	5 September 2015	1	1
Waikato	Mokau	961	-38.7	174.6	15 July 2015	2 December 2015	16	3
Waikato	Waikato	752	-37.3	174.8	15 August 2015	23 November 2015	13	3
Waikato	Awakino	428	-38.7	174.8	17 July 2015	5 December 2015	10	2
Waikato	Waikawau	282	-38.5	174.8	24 September 2015	26 November 2015	5	2
Waikato	Waingaro	141	-37.7	175.0	2 October 2015	27 November 2015	4	1
Waikato	Marokopa	197	-38.3	174.7	23 September 2015	2 November 2015	3	1
Waikato	Wentworth	241	-37.2	175.9	7 October 2015	15 November 2015	3	1
Waikato	Oparau	50	-38.1	174.9	1 October 2015	1 October 2015	1	1
Wairarapa	Whareama	40	-41.0	176.1	31 August 2015	31 August 2015	1	1
Wellington	Otaki	35	-40.8	175.1	12 October 2015	26 October 2015	2	1
Wellington	Pauatahanui	102	-41.1	174.9	14 October 2015	24 October 2015	2	1
Wellington	Waikanae	82	-40.9	175.0	30 September 2015	3 October 2015	2	1
Wellington	Hutt	82	-41.2	174.9	30 September 2015	30 September 2015	1	1
Wellington	LakeFerry	44	-41.4	175.1	30 November 2015	30 November 2015	1	1
Wellington	PekaPeka	67	-40.8	175.1	18 October 2015	18 October 2015	1	1
Westland	Hokitika	564	-42.7	171.0	6 July 2015	8 December 2015	8	3
Westland	Wanganui	344	-43.0	170.4	7 July 2015	1 December 2015	8	3
Westland	Waiatoto	1168	-44.0	168.8	8 August 2015	18 December 2015	15	2
Westland	Waimea	294	-42.6	171.1	24 July 2015	7 December 2015	9	2
Westland	Paringa	148	-43.6	169.4	16 September 2015	4 November 2015	3	2
Westland	Cascade	410	-44.0	168.4	8 September 2015	13 November 2015	6	1
Westland	Okarito	135	-43.2	170.2	16 September 2015	11 November 2015	3	1
Westland	Taramakau	61	-42.6	171.1	14 September 2015	14 September 2015	1	1

Appendix 2: Missing Values

Table A2: The proportion of missing values from each morphometric measure by regions, sampler and species. The highest proportion of missing values is depth measures from six samplers where there were no depths measures taken from some species of fish in their samples.

Region	Sampler	Species	n	Proportion Missing		
				Depth	Length	Weight
Wellington	David	koaro	1	0.0000	1.0000	0.0000
Tasman	Gary	kokopuBand	3	0.0000	1.0000	0.0000
Waikato	Marie	koaro	3	0.0000	0.6667	0.0000
Waikato	Marie	kokopuBand	9	0.1111	0.4444	0.1111
Buller	Eimear	inanga	22	0.0000	0.4091	0.0455
HawkesBay	Matt	inanga	29	0.0690	0.3793	0.0000
HawkesBay	Alans Son	kokopuBand	3	0.0000	0.3333	0.0000
Canterbury	Mike	inanga	100	0.6000	0.3100	0.3000
Otago	Peter	koaro	10	0.0000	0.3000	0.0000
Waikato	Bryan	kokopuBand	15	0.0000	0.2667	0.0000
Westland	Mark/Karen	inanga	4	0.0000	0.2500	0.0000
Southland	Jan	inanga	321	0.0000	0.2243	0.2150
Canterbury	John	inanga	149	0.2081	0.2148	0.2013
Waikato	Kevin	koaro	5	0.6000	0.2000	0.0000
HawkesBay	Karena	kokopuBand	10	0.0000	0.2000	0.0000
Tasman	Gary	inanga	21	0.0000	0.1905	0.0952
Southland	Harry	kokopuBand	16	0.0000	0.1875	0.0000
Marlborough	George	kokopuBand	11	0.0000	0.1818	0.0000
Waikato	Bryan	koaro	6	0.0000	0.1667	0.0000
Buller	Bruce	kokopuBand	6	0.0000	0.1667	0.0000
Westland	Gary	kokopuBand	43	0.4186	0.1628	0.0465
Westland	Gary	inanga	276	0.5326	0.1449	0.1413
BOP	Hine	kokopuBand	21	0.0476	0.1429	0.0476
Wellington	David	kokopuBand	28	0.0000	0.1429	0.0000
Waikato	Jenny	kokopuBand	22	0.0909	0.1364	0.0000
Waikato	Jenny	inanga	150	0.2000	0.1267	0.0000
Westland	Graham/Brenda	kokopuBand	8	0.0000	0.1250	0.1250
Otago	Mark	koaro	8	0.0000	0.1250	0.0000
Buller	Tony	kokopuBand	8	0.0000	0.1250	0.0000
Waikato	Kevin	kokopuBand	50	0.1200	0.1200	0.0000
Buller	Tony	koaro	53	0.0377	0.1132	0.0000
Waikato	Neville	kokopuBand	9	0.0000	0.1111	0.0000
Tasman	Gary	koaro	9	0.0000	0.1111	0.0000
Southland	Harry	koaro	49	0.0000	0.1020	0.0000
Wellington	Kris	kokopuBand	10	0.0000	0.1000	0.0000
Otago	Jo	kokopuBand	10	0.0000	0.1000	0.0000
HawkesBay	Matt/Dan	inanga	40	0.0000	0.1000	0.0000

Buller	Bearill	inanga	40	0.0000	0.1000	0.0000
BOP	Kelly	inanga	220	0.0000	0.1000	0.0182
Westland	Neville	koaro	41	0.0000	0.0976	0.0000
Buller	Tony	inanga	52	0.0000	0.0962	0.0000
Waikato	Ralf	kokopuBand	67	0.0149	0.0896	0.0149
Westland	Gary	koaro	91	0.6593	0.0879	0.0549
Westland	Mike	inanga	46	0.0000	0.0870	0.0000
Waikato	Shelley	kokopuBand	35	0.0000	0.0857	0.0000
HawkesBay	Alan	kokopuBand	12	0.0833	0.0833	0.0000
BOP	Tio	kokopuBand	12	0.0000	0.0833	0.0000
Waikato	Kevin	inanga	350	0.2000	0.0714	0.0000
Canterbury	Peter	kokopuBand	14	0.0000	0.0714	0.0000
Buller	Chrissy	inanga	89	0.0000	0.0674	0.0000
Buller	Ross	kokopuBand	104	0.0192	0.0673	0.0192
Westland	Des	inanga	194	0.0361	0.0670	0.0052
Wellington	David	inanga	120	0.0000	0.0667	0.0083
Marlborough	Tim	koaro	16	0.0000	0.0625	0.0000
Westland	Des	kokopuBand	50	0.0200	0.0600	0.0400
Waikato	Ralf	koaro	18	0.0000	0.0556	0.0000
Taranaki	Win	inanga	40	0.0000	0.0500	0.0000
Otago	Mark	inanga	40	0.0000	0.0500	0.0000
HawkesBay	Alans Son	inanga	40	0.0000	0.0500	0.0000
HawkesBay	Jeff	kokopuBand	20	0.0000	0.0500	0.0000
BOP	Wayne	inanga	40	0.0000	0.0500	0.0000
Buller	Ross	koaro	61	0.0000	0.0492	0.0000
HawkesBay	Karena	inanga	84	0.0000	0.0476	0.0000
Waikato	Marie	inanga	290	0.1379	0.0448	0.0000
Westland	Fay	inanga	415	0.0000	0.0434	0.0000
Westland	Des	kokopuGiant	48	0.0000	0.0417	0.0208
Westland	Raewyn	inanga	120	0.0000	0.0417	0.0000
Waikato	Shelley	inanga	120	0.0000	0.0417	0.0000
HawkesBay	Alan	inanga	120	0.0000	0.0417	0.0000
Westland	Fay	kokopuBand	75	0.0267	0.0400	0.0000
Tasman	Barbara	kokopuBand	25	0.0000	0.0400	0.0000
Waikato	Shelley	koaro	26	0.0000	0.0385	0.0000
Otago	Jo	inanga	52	0.0000	0.0385	0.0000
BOP	Kui	inanga	80	0.0000	0.0375	0.0000
Wellington	Alby	inanga	28	0.0357	0.0357	0.0357
Waikato	William	kokopuBand	122	0.0000	0.0328	0.0000
Waikato	Eddie	inanga	132	0.6970	0.0303	0.0000
Westland	Mark	inanga	104	0.0000	0.0288	0.0000
Buller	Ross	inanga	142	0.0070	0.0282	0.0070
Otago	Athol	kokopuBand	38	0.0263	0.0263	0.0000
Westland	Neville	inanga	240	0.0000	0.0250	0.0000

Waikato	Bryan	inanga	80	0.0000	0.0250	0.0000
Waikato	Ralf	inanga	240	0.0000	0.0250	0.0000
Taranaki	Dennis	kokopuBand	40	0.0000	0.0250	0.0000
Taranaki	Win	kokopuBand	40	0.0000	0.0250	0.0000
Southland	Harry	inanga	80	0.0000	0.0250	0.0000
Otago	Michael	inanga	80	0.0000	0.0250	0.0000
Manawatu	Beryl	inanga	80	0.0000	0.0250	0.0000
Canterbury	Mark	inanga	40	0.0000	0.0250	0.2500
Buller	Bruce	inanga	40	0.0000	0.0250	0.0000
Canterbury	Eimear	inanga	41	0.0244	0.0244	0.0000
Buller	Mike	kokopuBand	41	0.0000	0.0244	0.0000
Buller	Mike	inanga	90	0.5556	0.0222	0.0000
Buller	Warren	inanga	159	0.0000	0.0189	0.0000
Buller	John	inanga	54	0.0000	0.0185	0.0000
Buller	John	kokopuBand	58	0.0000	0.0172	0.0000
Waikato	William	inanga	119	0.3361	0.0168	0.0000
Marlborough	George	inanga	120	0.0000	0.0167	0.0000
Manawatu	Jim	kokopuBand	65	0.0000	0.0154	0.0000
Tasman	Sean	kokopuBand	305	0.0000	0.0131	0.0033
Westland	Graham/Brenda	inanga	80	0.0000	0.0125	0.0000
Tasman	Sean	inanga	649	0.0015	0.0108	0.0000
Buller	Warren	koaro	95	0.0000	0.0105	0.0000
Southland	Elaine	inanga	199	0.5980	0.0101	0.0000
Buller	Chrissy	koaro	200	0.0000	0.0100	0.0000
Buller	Mark	koaro	102	0.0882	0.0098	0.0098
Westland	Fay	koaro	211	0.0000	0.0095	0.0000
Southland	Kim	koaro	112	0.0000	0.0089	0.0000
Westland	Simon	inanga	240	0.0000	0.0083	0.0375
Tasman	Sean	koaro	275	0.0000	0.0073	0.0000
Manawatu	Lindsay	inanga	320	0.0000	0.0063	0.0000
Southland	Kim	inanga	360	0.0000	0.0056	0.0000
BOP	Peter	inanga	593	0.5413	0.0034	0.0000
Waikato	Eddie	kokopuBand	9	1.0000	0.0000	0.0000
Manawatu	Beryl	kokopuBand	1	1.0000	0.0000	0.0000
Manawatu	Beryl	koaro	1	1.0000	0.0000	0.0000
Manawatu	John	kokopuGiant	2	1.0000	0.0000	0.0000
Auckland	Other whitebaiters (written as Kim)	kokopuBand	30	1.0000	0.0000	0.0000
Auckland	Other whitebaiters (written as Kim)	inanga	30	1.0000	0.0000	0.0000
Buller	Mark	kokopuBand	49	0.6122	0.0000	0.0000
Southland	Brett	inanga	80	0.5000	0.0000	0.0000
BOP	Peter	koaro	2	0.5000	0.0000	0.0000
Waikato	John	inanga	452	0.4690	0.0000	0.0000
Waikato	John	koaro	4	0.2500	0.0000	0.0000

Westland	Peter	koaro	190	0.2105	0.0000	0.0000
Westland	Peter	inanga	240	0.1667	0.0000	0.0000
Canterbury	Ricky	koaro	6	0.1667	0.0000	0.0000
Canterbury	Ricky	inanga	194	0.1546	0.0000	0.0000
Waikato	Colleen	inanga	280	0.1429	0.0000	0.0000
Waikato	John	kokopuBand	48	0.1250	0.0000	0.0000
BOP	Peter	kokopuBand	17	0.1176	0.0000	0.0000
Canterbury	Fiona	inanga	270	0.1148	0.0000	0.0000
Canterbury	Fiona	kokopuBand	11	0.0909	0.0000	0.0000
Westland	Peter	kokopuBand	35	0.0857	0.0000	0.0000
Wellington	Darron	inanga	40	0.0250	0.0000	0.0000
Waikato	Peter	inanga	120	0.0083	0.0000	0.0000
Westland	Des	koaro	1	0.0000	0.0000	0.0000
Westland	Fay	kokopuGiant	1	0.0000	0.0000	0.0000
Westland	Graham	kokopuBand	1	0.0000	0.0000	0.0000
Westland	Graham	koaro	7	0.0000	0.0000	0.0000
Westland	Graham	inanga	40	0.0000	0.0000	0.0000
Westland	Graham/Brenda	koaro	12	0.0000	0.0000	0.0000
Westland	Mark	kokopuGiant	3	0.0000	0.0000	0.0000
Westland	Mark	kokopuBand	1	0.0000	0.0000	0.0000
Westland	Mark Connors	koaro	21	0.0000	0.0000	0.0000
Westland	Mark Connors	inanga	40	0.0000	0.0000	0.0000
Westland	Neville	kokopuBand	15	0.0000	0.0000	0.0000
Westland	Peter	kokopuGiant	1	0.0000	0.0000	0.0000
Westland	Raewyn	kokopuGiant	1	0.0000	0.0000	0.0000
Westland	Raewyn	kokopuBand	14	0.0000	0.0000	0.0000
Westland	Simon	kokopuGiant	9	0.0000	0.0000	0.0000
Westland	Simon	kokopuBand	74	0.0000	0.0000	0.0000
Westland	Simon	koaro	132	0.0000	0.0000	0.0000
Wellington	Alby	koaro	7	0.0000	0.0000	0.0000
Wellington	Darron	kokopuGiant	7	0.0000	0.0000	0.0000
Wellington	Darron	kokopuBand	40	0.0000	0.0000	0.0000
Wellington	Darron	koaro	15	0.0000	0.0000	0.0000
Wellington	Khan/Peter	kokopuBand	2	0.0000	0.0000	0.0000
Wellington	Khan/Peter	koaro	2	0.0000	0.0000	0.0000
Wellington	Khan/Peter	inanga	40	0.0000	0.0000	0.0000
Wellington	Kris	koaro	32	0.0000	0.0000	0.0000
Wellington	Kris	inanga	40	0.0000	0.0000	0.0000
Wairarapa	Graham	inanga	40	0.0000	0.0000	0.0000
Waikato	Colleen	kokopuGiant	3	0.0000	0.0000	0.0000
Waikato	Colleen	kokopuBand	123	0.0000	0.0000	0.0000
Waikato	Colleen	koaro	10	0.0000	0.0000	0.0000
Waikato	Jenny	kokopuGiant	1	0.0000	0.0000	0.0000
Waikato	Jenny	koaro	3	0.0000	0.0000	0.0000

Waikato	John	kokopuGiant	1	0.0000	0.0000	0.0000
Waikato	Kevin	kokopuGiant	1	0.0000	0.0000	0.0000
Waikato	Neville	koaro	1	0.0000	0.0000	0.0000
Waikato	Neville	inanga	40	0.0000	0.0000	0.0000
Waikato	Pete	inanga	8	0.0000	0.0000	0.0000
Waikato	Peter	kokopuBand	66	0.0000	0.0000	0.0000
Waikato	Peter	koaro	11	0.0000	0.0000	0.0000
Waikato	Ralf	kokopuGiant	3	0.0000	0.0000	0.0000
Tasman	Barbara	koaro	40	0.0000	0.0000	0.0000
Tasman	Barbara	inanga	40	0.0000	0.0000	0.0000
Tasman	Peter	inanga	40	0.0000	0.0000	0.0000
Tasman	Sean	kokopuGiant	4	0.0000	0.0000	0.0000
Taranaki	Dennis	koaro	9	0.0000	0.0000	0.0000
Taranaki	Dennis	inanga	40	0.0000	0.0000	0.0000
Taranaki	Diane	kokopuBand	1	0.0000	0.0000	0.0000
Taranaki	Diane	koaro	10	0.0000	0.0000	0.0000
Taranaki	Diane	inanga	40	0.0000	0.0000	0.0000
Taranaki	Tony	kokopuBand	40	0.0000	0.0000	0.0000
Taranaki	Tony	inanga	40	0.0000	0.0000	0.0000
Taranaki	Win	koaro	1	0.0000	0.0000	0.0000
Southland	A.L. McDonald	kokopuBand	16	0.0000	0.0000	0.0000
Southland	A.L. McDonald	koaro	1	0.0000	0.0000	0.0000
Southland	A.L. McDonald	inanga	40	0.0000	0.0000	0.0000
Southland	Elaine	koaro	7	0.0000	0.0000	0.0000
Southland	Jan	kokopuBand	1	0.0000	0.0000	0.0000
Southland	Jan	koaro	3	0.0000	0.0000	0.0000
Southland	Kim	kokopuBand	26	0.0000	0.0000	0.0000
Southland	Robert	kokopuBand	1	0.0000	0.0000	0.0000
Southland	Robert	koaro	9	0.0000	0.0000	0.0000
Southland	Steve	inanga	120	0.0000	0.0000	0.0000
Otago	Athol	koaro	10	0.0000	0.0000	0.0000
Otago	Athol	inanga	80	0.0000	0.0000	0.0000
Otago	Brent	kokopuBand	5	0.0000	0.0000	0.0000
Otago	Brent	inanga	80	0.0000	0.0000	0.0000
Otago	Glenys/Ian	inanga	40	0.0000	0.0000	0.0000
Otago	Jo	koaro	1	0.0000	0.0000	0.0000
Otago	Michael	koaro	6	0.0000	0.0000	0.0000
Otago	Peter	kokopuBand	8	0.0000	0.0000	0.0000
Otago	Peter	inanga	80	0.0000	0.0000	0.0000
Marlborough	George	koaro	4	0.0000	0.0000	0.0000
Marlborough	Ken	kokopuBand	2	0.0000	0.0000	0.0000
Marlborough	Ken	inanga	40	0.0000	0.0000	0.0000
Marlborough	Tim	kokopuBand	13	0.0000	0.0000	0.0000
Marlborough	Tim	inanga	120	0.0000	0.0000	0.0000

Manawatu	Deb	inanga	40	0.0000	0.0000	0.0000
Manawatu	Eimear	kokopuBand	3	0.0000	0.0000	0.0000
Manawatu	Eimear	inanga	40	0.0000	0.0000	0.0000
Manawatu	Jim	koaro	21	0.0000	0.0000	0.0000
Manawatu	Jim	inanga	80	0.0000	0.0000	0.0000
Manawatu	John	kokopuBand	7	0.0000	0.0000	0.0000
Manawatu	John	koaro	3	0.0000	0.0000	0.0000
Manawatu	John	inanga	40	0.0000	0.0000	0.0000
Manawatu	Lindsay	kokopuGiant	6	0.0000	0.0000	0.0000
Manawatu	Lindsay	kokopuBand	46	0.0000	0.0000	0.0000
Manawatu	Lindsay	koaro	92	0.0000	0.0000	0.0000
HawkesBay	Alan	koaro	16	0.0000	0.0000	0.0000
HawkesBay	Alans Son	koaro	5	0.0000	0.0000	0.0000
HawkesBay	Dan	kokopuBand	2	0.0000	0.0000	0.0000
HawkesBay	Dan	koaro	10	0.0000	0.0000	0.0000
HawkesBay	Dan	inanga	200	0.0000	0.0000	0.0000
HawkesBay	Jeff	koaro	21	0.0000	0.0000	0.0000
HawkesBay	Jeff	inanga	161	0.0000	0.0000	0.0000
HawkesBay	Karena	koaro	1	0.0000	0.0000	0.0000
HawkesBay	Roger	kokopuBand	1	0.0000	0.0000	0.0000
HawkesBay	Roger	inanga	40	0.0000	0.0000	0.0000
HawkesBay	Tyrone	inanga	40	0.0000	0.0000	0.0000
Canterbury	Colin	kokopuBand	6	0.0000	0.0000	0.0000
Canterbury	Colin	inanga	40	0.0000	0.0000	0.0000
Canterbury	Desmond	inanga	40	0.0000	0.0000	0.0000
Canterbury	Fiona	koaro	12	0.0000	0.0000	0.0000
Canterbury	John	kokopuBand	13	0.0000	0.0000	0.0000
Canterbury	Kerry	inanga	40	0.0000	0.0000	0.0000
Canterbury	Peter	koaro	43	0.0000	0.0000	0.0000
Canterbury	Peter	inanga	40	0.0000	0.0000	0.0000
Canterbury	Ricky	kokopuBand	3	0.0000	0.0000	0.0000
Canterbury	Russell	kokopuBand	3	0.0000	0.0000	0.0000
Canterbury	Russell	inanga	40	0.0000	0.0000	0.0000
Canterbury	Samson	inanga	40	0.0000	0.0000	0.0000
Canterbury	Steve	kokopuBand	4	0.0000	0.0000	0.0000
Canterbury	Steve	inanga	320	0.0000	0.0000	0.0000
Canterbury	Val	inanga	40	0.0000	0.0000	0.0000
Canterbury	Willy	inanga	80	0.0000	0.0000	0.0000
Buller	Bearill	kokopuGiant	3	0.0000	0.0000	0.0000
Buller	Bearill	kokopuBand	1	0.0000	0.0000	0.0000
Buller	Chrissy	kokopuBand	161	0.0000	0.0000	0.0000
Buller	John	koaro	25	0.0000	0.0000	0.0000
Buller	Mark	kokopuGiant	13	0.0000	0.0000	0.0000
Buller	Mark	inanga	120	0.0000	0.0000	0.0000

Buller	Mike	koaro	44	0.0000	0.0000	0.0000
Buller	Pauline	kokopuGiant	16	0.0000	0.0000	0.0000
Buller	Pauline	kokopuBand	3	0.0000	0.0000	0.0000
Buller	Pauline	inanga	40	0.0000	0.0000	0.0000
Buller	Selwyn	kokopuBand	4	0.0000	0.0000	0.0000
Buller	Selwyn	inanga	40	0.0000	0.0000	0.0000
Buller	Warren	kokopuGiant	2	0.0000	0.0000	0.0000
Buller	Warren	kokopuBand	54	0.0000	0.0000	0.0000
BOP	Beryls	kokopuBand	31	0.0000	0.0000	0.0000
BOP	Beryls	koaro	11	0.0000	0.0000	0.0000
BOP	Beryls	inanga	40	0.0000	0.0000	0.0000
BOP	Brian	kokopuBand	2	0.0000	0.0000	0.0000
BOP	Brian	koaro	17	0.0000	0.0000	0.0000
BOP	Brian	inanga	80	0.0000	0.0000	0.0000
BOP	Estelle	kokopuBand	1	0.0000	0.0000	0.0000
BOP	Estelle	inanga	1	0.0000	0.0000	0.0000
BOP	Hine	koaro	1	0.0000	0.0000	0.0000
BOP	Hine	inanga	80	0.0000	0.0000	0.0000
BOP	Kelly	kokopuBand	2	0.0000	0.0000	0.0000
BOP	Kelly	koaro	20	0.0000	0.0000	0.0000
BOP	Kerry	kokopuBand	40	0.0000	0.0000	0.0000
BOP	Kerry	koaro	1	0.0000	0.0000	0.0000
BOP	Kerry	inanga	40	0.0000	0.0000	0.0000
BOP	Kui	kokopuBand	26	0.0000	0.0000	0.0000
BOP	Kui	koaro	2	0.0000	0.0000	0.0000
BOP	Tio	inanga	80	0.0000	0.0000	0.0000
Auckland	Kim	kokopuBand	8	0.0000	0.0000	0.0000
Auckland	Kim	inanga	80	0.0000	0.0000	0.0000